Final Assignment

Psychological assessment guides are created by psychology professionals to provide the public with accurate and authoritative information appropriate for their current needs. Information available to the public about psychological testing and assessment varies widely depending on the professional creating it, the purpose of the assessment, and the intended audience. When professionals effectively educate the public on the *how*, *what*, and *why* behind assessments and the strengths and limitations of commonly used instruments, potential clients are in a better position to be informed users of assessment products and services. The Assessment Guides developed in this course will be designed to provide the lay public with accurate and culturally relevant information to aid them in making informed decisions about psychological testing. Students will develop their Guides with the goal of educating readers to be informed participants in the assessment process.

There is no required template for the development of the Assessment Guide. Students are encouraged to be creative while maintaining the professional appearance of their work. The Guide must be reader-friendly (sixth- to ninth-grade reading level) and easy to navigate, and it must include a combination of text, images, and graphics to engage readers in the information provided. Throughout their Guides, students will provide useful examples and definitions as well as questions readers should ask their practitioners. To ensure accuracy, students are expected to use only scholarly and peer-reviewed sources for the information in the development of their Guides.

Students will begin their Guides with a general overview of assessment, reasons for assessment referrals, and the importance of the role of each individual in the process. Within each of the remaining sections, students will describe the types of assessments that their readers may encounter, the purposes of each type of assessment, the different skills and abilities the instruments measure, the most valid and reliable uses of the measures, and limitations of the measures. A brief section will be included to describe the assessment process, the types of professionals who conduct the assessments, and what to expect during the assessment meetings.

The Assessment Guide must include the following sections:

Table of Contents (Portrait orientation must be used for the page layout of this section.)
In this one-page section, students must list the following subsections and categories of assessments.

- Introduction and Overview
- Tests of Intelligence
- Tests of Achievement
- Tests of Ability
- Neuropsychological Testing
- Personality Testing
- Industrial, Occupational, and Career Assessment
- Forensic Assessment
- Special Topics (student's choice)

- References

Section 1: Introduction and Overview (Portrait or landscape orientation may be used for the page layout of this section.)
Students will begin their Guides with a general overview of assessment. In this two-page section, students will briefly address the major aspects of the assessment process. Students are encouraged to develop creative titles for these topics that effectively communicate the meanings to the intended audience.

- Definition of a Test (e.g., What is a Test?)
- Briefly define psychological assessment.
- Types of Tests
- Identify the major categories of psychological assessment.
- Reliability and Validity
- Briefly define the concepts of reliability and validity as they apply to psychological assessment.
- Role of testing and assessment in the diagnostic process
- Briefly explain role of assessment in diagnosis.
- Professionals Who Administer Tests
- Briefly describe the types of professionals involved in various assessment processes.
- Culture and Testing
- Briefly describe issues of cultural diversity as it applies to psychological assessment.

Categories of Assessment (Portrait or landscape orientation may be used for the page layout of this section.)
For each of the following, students will create a two-page information sheet or pamphlet to be included in the Assessment Guide. For each category of assessment, students will include the required content listed in the PSY640 Content for Testing Pamphlets and Information Sheets (Links to an external site.). Be sure to reference the content requirements (Links to an external site.) prior to completing each of the information sheets on the following categories of assessment.

- Tests of Intelligence
- Tests of Achievement
- Tests of Ability
- Neuropsychological Testing
- Personality Testing
- Industrial, Occupational, and Career Assessment
- Forensic Assessment
- Special Topics (Students will specify which topic they selected for this pamphlet or information sheet. Additional instructions are noted below.)

Special Topics (Student's Choice)
In addition to the required seven categories of assessment listed above, students will develop an eighth information sheet or pamphlet that includes information targeted

either at a specific population or about a specific issue related to psychological assessment not covered in one of the previous sections. Students may choose from one of the following categories:

- Testing Preschool-Aged Children
- Testing Elementary School-Aged Children
- Testing Adolescents
- Testing Geriatric Patients
- Testing First Generation Immigrants
- Testing in Rural Communities
- Testing English Language Learners
- Testing Individuals Who Are (Select one: Deaf, Blind, Quadriplegic)
- Testing Individuals Who Are Incarcerated
- Testing for Competency to Stand Trial
- Testing in Child Custody Cases

References (Portrait orientation must be used for the page layout of this section.) Include a separate reference section that is formatted according to APA style *as outlined in the* Ashford Writing Center (Links to an external site.). The reference list must consist entirely of scholarly sources. For the purposes of this assignment, assessment manuals, the course textbook, chapters from graduate-level textbooks, chapters from professional books, and peer-reviewed journal articles may be used as resource material. A minimum of 16 unique scholarly sources including a minimum of 12 peer-reviewed articles published within the last 10 years from the Ashford University Library must be used within the Assessment Guide. The bulleted list of credible professional and/or educational online resources required for each assessment area will not count toward these totals.

Attention Students: The Masters of Arts in Psychology program is utilizing the Pathbrite portfolio tool as a repository for student scholarly work in the form of signature assignments completed within the program. After receiving feedback for this Assessment Guide, please implement any changes recommended by the instructor, go to Pathbrite (Links to an external site.)and upload the revised Assessment Guide to the portfolio. (Use the Pathbrite Quick-Start Guide (Links to an external site.) to create an account if you do not already have one.) The upload of signature assignments will take place after completing each course. Be certain to upload revised signature assignments throughout the program as the portfolio and its contents will be used in other courses and may be used by individual students as a professional resource tool. See the Pathbrite (Links to an external site.) website for information and further instructions on using this portfolio tool.

The Assessment Guide

- Must be 18 pages in length (not including title and reference pages) and formatted according to APA style as outlined in the Ashford Writing Center (Links to an external site.).
- Must include a separate title page with the following:
  - Title of guide

  - Student's name
  - Course name and number
  - Instructor's name
  - Date submitted
- Must use at least 16 scholarly sources, including a minimum of 12 peer-reviewed articles from the Ashford University Library.
- Must document all sources in APA style as outlined in the Ashford Writing Center.
- Must include a separate reference page that is formatted according to APA style as outlined in the Ashford Writing Center.
- Must incorporate at least three different methods of presenting information (e.g., text, graphics, images, original cartoons

PSY640 Content for Testing Pamphlets and Information Sheets

For each category of assessment listed in the assignment, students will create two pages of information. The intent for the layout is that it be consistent with either a two-page information sheet (front and back), or a two-sided tri-fold pamphlet that might be found in the office of a mental health professional. The presentation of the information within each pamphlet or brochure must incorporate at least three different visual representations of the information (e.g., text, graphics, images, original cartoons).

For each pamphlet or information sheet a minimum of three scholarly sources must be used, at least two of which must be from peer-reviewed journal articles published within the last 10 years and obtained from the Ashford University Library. Some sources may be relevant for more than one category of assessment; therefore, it is acceptable to use relevant sources in more than one category. Remember that the language for each information sheet should be at the sixth- to ninth-grade reading level to allow a broad audience at various ages and levels of education to better understand each category of assessment.

For each category of assessment.

- Introduce and offer a brief, easy-to-understand definition for the broad assessment category being measured. (e.g., What is intelligence?, What is achievement?, What is personality?, What does "neuropsychological" mean? What does "forensic" mean?)
- Provide a brief overview of the types of tests commonly used within the category of assessment explain what they measure. Compare the commonly used assessment instruments within the category.
- Describe appropriate and inappropriate uses of tests within the category of assessment. Explain why some tests are more appropriate for specific populations and purposes and which tests may be inappropriate. Analyze and describe the challenges related to assessing individuals from diverse social and cultural backgrounds. Evaluate the ethical interpretation of testing and assessment data as it relates to the test types within the category. Describe major debates in the field regarding different assessment approaches within the category. (e.g., Intellectual disabilities, formerly known as "mental retardation," cannot be determined by a single test. Thus, an inappropriate use of an intelligence test would be to use such a test as the sole instrument to diagnose an intellectual ability.)
- Describe the format in which assessment results may be expected. Evaluate and explain the professional interpretation of testing and assessment data. Analyze the psychometric methodologies typically employed in the validation of types of psychological testing within the

category. Include information about the types of scores used to communicate assessment results consistent with the tests being discussed (e.g., scaled scores, percentile rank, grade equivalent, age equivalent, standard age score, confidence interval).

- Explain the common terminology used in assessment in a manner that demystifies the professional jargon (e.g., In the course of discussing intelligence testing, students would define concepts such as I.Q., categories of intelligence, and the classification labels used to describe persons with intellectual disabilities.)
- Include a bulleted list of at least three credible professional and/or educational online resources where the reader can obtain more information about the various types of testing in order to aid him or her in the evaluation and interpretation of testing and assessment data. No commercial websites may be used. Include the name of the organization that authored the web page, the title of the web page and/or document, and the URL. (These websites will not count toward the 12 scholarly resources required for the assignment.)

JOURNAL OF PERSONALITY ASSESSMENT, *83*(2), 120–127

STATISTICAL DEVELOPMENTS AND APPLICATIONS
SPECIAL SERIES: Personality Measurement

Personality in Proportion: A Bipolar Proportional Scale for Personality Assessments and Its Consequences for Trait Structure

Willem K. B. Hofstee and Jos M. F. Ten Berge
The Heymans Institute
University of Groningen

Trait structures resulting from personality assessments on Likert scales are affected by the additive and multiplicative transformations implied in interval scaling and correlational analysis. The effect comes into view on selecting a plausible alternative scale. To this end, we propose a bipolar bounded scale ranging from −1 to +1 representing an underlying process in which the assessor would review and discount positive and negative behavioral instances of a trait. As an appropriate index of likeness between variables X and Y, we propose $L_{XY} = \Sigma XY/N$, the average of the raw scores cross products. Using this index, we carried out a raw scores principal component analysis on data consisting of 133 participants who had each been rated by 5 assessors including self on 914 items. Contrary to the Big-Five structure that was found in these data on standard analysis, the results showed a relatively large first principal component F_1 and 2 very small ones, F_2 and F_3. The sizes $L_{FF} = \Sigma F^2/N$, the averages of the squared component scores, were modest to small. It thus appears that the scale, bipolar proportional versus standard, has a profound impact on the size and structure of personality assessments. The dissimilarity remains on analyzing self-ratings rather than averaged (over the 5 assessors) ratings.

We consider personality assessment as a way of communicating about people's traits or qualities. The main parties are the assessors and their audience, consisting of the target persons themselves or third parties. In between, there are investigators represented by instruments for data gathering—questionnaires in a wide sense—and psychometric and statistical procedures for data processing. In this perspective, their task is to facilitate the communication between assessor and audience by asking the right questions, by adequately summarizing the data, removing biases, and the like. We propose that standard procedures for data processing imported from the area of objective tests and measurement and consisting of relative scaling and correlational analysis may not be the most obvious choice from this communicative point of view. We present a bipolar proportional scale as an alternative. We give an empirical demonstration in which the set of alternative procedures is applied. It appears that they have profound implications for prevailing conceptions about the size of individual differences and their structure. We discuss whether these implications are nonetheless acceptable.

REPRESENTING ASSESSMENT DATA

In developing alternative procedures, we refer to the point of view of the assessor (and the audience). Evidently, that perspective represents a construction. We do not claim to be able to look inside the heads of assessors; what we argue is that the construction is recognizable and plausible. We also do not assume that the assessor is necessarily right; we analyze whether the reconstruction is rational. Most notably, we do not assert that relative scaling is wrong. In comparative contexts, for example, in strictly comparative personnel selection, relative models may be preferable over this one.

A Bipolar Scale

The pivot of our assessment model is a conception of the midpoint of a Likert scale (e.g., 3 on a scale ranging from 1 to 5) as a natural zero point ruling out additive transformations. Personality traits or qualities in general are understood in a bipolar fashion: Negations (e.g., unfriendly) do not denote the mere

absence of a trait but its reverse. Opposites have high negative correlations after removal of acquiescence (Hofstee, Ten Berge, & Hendriks, 1998); traits such as unfriendly are rated socially undesirable rather than neutral as would logically be the case with null friendliness. Such negations are thus understood in the manner of a *litotes* (expressing of an affirmative by the negative of its contrary). In the typical case, most people are assessed at the socially desirable side of the midpoint of the scale. Therefore, for persons with a score between the midpoint of the scale and the mean of the distribution, relative scaling would reverse the sign of the assessment. For the others, the shift is less dramatic but still substantial.

One could argue that relative scaling corrects for a bias, namely, socially desirable responding, instead of introducing one. That position, however, is difficult to maintain. First, social desirability is not confined to self-report: On average, all sorts of assessors judge people to be socially desirable, although some are judged more desirable than others, and although assessors differ to some extent. Second, assessors including self agree on the differential social desirability of target individuals, making social desirability a trait (a target person effect) rather than a mere response style (an assessor effect). Thus, an absolute conception of the scale midpoint provides a plausible representation of the assessor's point of view.

We do not presume that personality traits are bipolar, for example, at a behavioral level. Whereas an individual can meaningfully be assessed to be unforgiving—meaning the opposite of forgiving—it may not be so easy to "unforgive" someone whether linguistically or behaviorally. Nonetheless, behavioral and verbal specifications of traits, such as respects others versus looks down on others, appear to show the same bipolar structure as trait adjectives (the example was taken from Hendriks's, 1997, p. 121–122, study, in which these items had high opposite loadings on one and the same factor).

Bounded Scales

Hofstee and Hendriks (1998), on the basis of the preceding argument, proposed the use of scales anchored at the scale midpoint rather than the mean of the distribution. In their procedure, the spread of the scores is set at unity by dividing the deviation scores (from the midpoint instead of the mean) by their standard deviation. The procedure has the advantage of leaving correlational and factor structures untouched. Here, we propose a bounded scale that may be found to come closer to the assessor's perspective.

All data-gathering procedures use bounded scales. One could think of an unbounded scale: Assessors might be instructed to assign any number between plus and minus infinity to a person's friendliness. We do not know of any case in which such scales, whether bipolar or unipolar, have been applied. An obvious reason is that the assessor would have to decide whether John's utter friendliness should be rated at +1,000 or +100,000, which is difficult to do. A related reason is that the ratings of two assessors would be all but incomparable. The problem is not that the number of scale points is infinite: Likert scales may have large numbers of scale points up to infinity if the assessor is instructed to place a mark on a line (a procedure that might have become more popular with automatic computerized scoring). It is unbounded scales that are problematic. Moreover, observed distributions on different traits differ in spread. Relatively neutral traits may span the whole scale, but it would be difficult to find any case in which a person is assessed to be totally or extremely unreliable or murderous. In distributions with small standard deviations, as with clearly desirable or undesirable traits, assessments in the mildly undesirable region may easily become extreme (e.g., minus 3 *SD*s) on transformation. Presumably, that is not what the assessor had in mind; it is definitely not what he or she has said. In other words, the assumption underlying classical standardization—namely, that spreads are arbitrary and carry no meaning—is not fulfilled.

We propose to standardize assessments at the scale end, that is, adopting a bipolar proportion (or percentage) scale running from −1 to +1 (or −100 to +100). Scores X on Likert scales are linearly transformed by taking

$$(X - \tfrac{1}{2}h - \tfrac{1}{2}g)/(\tfrac{1}{2}h - \tfrac{1}{2}g), \quad (1)$$

with h the highest possible scale value and g the lowest. In simpler terms, set the scale midpoint at 0, the scale ends at −1 and +1, and interpolate linearly. For example

Rating on 5-point scale	1	2	3	4	5
Transformation	−1	−.5	0	+.5	+1

For another example, any binary scale transforms into [−1, +1].

The proposed convention suggests an underlying assessment model that is well in line with classical notions in personality assessment, namely, Bem and Allen's (1974) summary label interpretation of trait ascription and Buss and Craik's (1983) act frequency approach to personality. According to these notions, the assessor reviews relevant behavioral instances of a trait, for example, cases in which the person has helped others, cheered them up, expressed an interest in them, and the like, when assessing that person's friendliness. In view of the bipolar conception of traits, we add that counterinstances are also relevant to the assessor, for example, unfriendly behaviors such as short-changing others, making them feel uncomfortable, turning one's back on them, and the like.

Consequently, we interpret a score of 4 on a 5-point Likert scale (transformed into +.5 on the bipolar proportion scale with .75 representing the boundary between the scale points +.5 and +1 and .25 the boundary between 0 and +.5) along the following lines: In relevant situations, this individual has shown a clear preponderance of friendly over unfriendly be-

haviors, for example, between 7 versus 1 and 5 versus 3 if the number of situations is 8 so that the proportion of friendly minus the proportion of unfriendly behaviors is at most (7 – 1)/8 = .75 and at least (5 – 3)/8 = .25. Clearly, one should not postulate deliberate and meticulous counts and calculations in the mind of the assessor; all that can be asked for are rough intuitive estimates and a discounting of positive and negative instances subject to all sorts of error. However, the model is at least correct in the sense that it describes what one would like the assessor to do: review and discount concrete instances and counterinstances relevant to the question.

The bipolar proportion model is not restricted to direct trait ratings. At a more specific level, the same estimation process can be postulated. Helping others versus turning one's back on them constitutes a large set of more specific relevant behaviors in relevant situations. For example, turning on the favorite music of a person who is cleaning the room may be found helpful. Even assisting a person crossing the street versus not doing so has subspecifications according to the density of the traffic, the status or demeanor of that person, and so on and so forth.

To summarize the argument thus far, a bipolar proportion scale with a natural zero point in the middle provides a plausible reconstruction of the assessment process. As either additive or multiplicative transformations are arguably inappropriate, our proposal amounts to conceiving of assessment scales as (bipolar) absolute scales. We are aware that the bipolar proportion model may be refined, for example, by allowing differential weights for instances according to their relevance or prototypicality, by taking assessment biases into account such as acquiescent responding and individual differences in that respect (see Hofstee et al., 1998), by establishing the value of intermediate scale points in an empirical manner instead of interpolating linearly, and so on.

Likeness Coefficients

The next problem is choosing a coefficient of association for absolute scales. Hofstee (2002) proposed adopting the most elementary measure of likeness, namely, the averaged cross-product $L_{XY} = \Sigma XY/N$. With scores between –1 and +1, L_{XY} is conveniently bounded within those same limits. As a consequence, $L_{XX} = \Sigma X^2/N$, the extent to which X is like itself, is generally not unity but ≤ 1; $L_{XX} = 1$ only if all ratings are at the extreme scale ends. So L_{XX} functions as an index of saliency: It represents the size of the **X** vector.

The primary argument in favor of L_{XY} is its analogy to $r_{XY} = \Sigma z_X z_Y/N$ for likeness between variables scored on interval scales: Both are averaged cross-products, which is how correlations, interactions, and likenesses in general are represented. For absolute scales, cross-products of the raw scores rather than z scores should be averaged.

Zegers and Ten Berge (1985) presented the identity coefficient $e_{XY} = 2\,\Sigma XY/(\Sigma X^2 + \Sigma Y^2)$, thus $2L_{XY}/(L_{XX} + L_{YY})$, as an appropriate association coefficient for (unbounded) absolute scales. The denominator is needed to keep e_{XY} within bounds. As one might wish to disregard the fact that the bipolar proportion scale happens to be bounded and prefer the identity coefficient, we present a detailed comparison of L_{XY} and e_{XY}.

Unlike r_{XY}, both coefficients are defined at the level of a single pair of observations, permitting sayings such as "In my case, friendliness and socialness do not go together." Both coefficients can thus accommodate the notion of intraindividual structure at a particular point in time, that is, without intraindividual replication. Unlike e_{XY}, however, L_{XY} at the aggregate level is the mean of the individuals' coefficients. L_{XY} is thus perfectly separable in the sense that the likeness coefficient at the individual level stays the same after aggregation; with e_{XY}, the denominator changes on taking other cases into account.

Using L_{XY}, two individuals of equal but moderate friendliness are less alike in that respect than two individuals with pronounced friendliness: More extreme or salient scores obtain higher weights. For example, L_{XY} is higher for two individuals with scores +.2 and +.7 than for two individuals with scores +.2 and +.2. Using e_{XY}, such effects would be corrected on calculating individual coefficients; they would, however, return if the coefficient is calculated at the aggregate level and then split up, as the denominator is a constant at that level.

A dramatic difference arises on comparing the sizes of the two coefficients. In the general case, $L_{XY} < e_{XY}$ as ΣX^2 and ΣY^2 are smaller than N. Only if all scores are +1 or –1, as would automatically be the case with a binary scale transformed into the bipolar proportion scale, would the two coefficients be identical. They would degenerate into a coefficient proposed by Holley and Guilford (1964), among others, which takes the value of the diagonal proportions minus the off-diagonal ones in the fourfold table (proportion agreement p_A minus proportion disagreement, or $2p_A - 1$). In the general case, size or saliency influences the likeness coefficient, whereas the identity coefficient disregards those aspects. Using continuous scales, assessments of ±.5 would be typical rather than extreme ones, so the typical likeness coefficient would be about four times as small as the corresponding identity coefficient. Another consequence is that variables with relatively small sizes L_{XX} will have relatively little impact on the multivariate structure of sets of variables as in factor analysis. The question is how to appreciate these properties of the likeness coefficient.

From the point of view of representing the concept of likeness, the coefficient may be found quite defensible. Take an example in which two individuals are equally but only vaguely at the friendly side of the scale; say their scores are .1, indicating a 55 to 45 preponderance of friendly over unfriendly behaviors. One may envisage a thought experiment in which the two persons would have been observed to behave in a friendly or unfriendly manner in the same situations, creating a 2 × 2 table with 55–45 marginal proportions.

Given behavioral inconsistencies and observer error and the resulting low correlations between situations, the off-diagonal proportions would be sizeable and the overlap would be modest. Conversely, extreme marginals for either or both variables automatically create more overlap and higher likeness coefficients. So to the extent that likeness or association is based on overlap rather than statistical dependence, L_{XY} captures it.

We submit that assessors and their public overwhelmingly opt for overlap rather than statistical dependence in defining likeness or even correlation. We base our prediction on informal classroom experiments that are easily replicated. Present people with fourfold tables with extreme marginal frequencies, for example, a table with 90 and 0 in the diagonal cells, and 5 in both off-diagonals (so that $r_{XY}=-.05$ and $L_{XY}=+.80$), and they will say that the two variables are clearly positively related. It does not even help much if the respondents have followed a course in applied statistics. One may of course object that they are mistaken and that coefficients that reflect overlap are wrong because they need not be zero on statistical independence. However, that argument is not nearly as straightforward as it may look. The layperson's point of view receives support from a classical argument mostly referred to as Meehl's paradox (Meehl & Rosen, 1955).

Take a diagnostic setting (rather than a comparative selection setting) in which clients or students are assessed on some trait or quality with a heavily skewed base rate, say 95% positive and 5% negative according to some reasonable criterion. Take two assessors, one of whom displays a 95–5 selection rate but a maximally negative validity ($r=-.05$, see previously), whereas the other has a 50–50 selection rate but maximally positive ($r = .23$) validity. By the rational standard consisting of proportions of correct diagnoses (90% vs. 55%), the first assessor outperforms the second. The difference is reflected in the values of L_{XY}, which are .80 and .10, respectively. Surely, a random procedure with a 95–5 selection rate would work even better than the first assessor, but that is because the procedure has been fed with prior knowledge: It has been told to take the base rate as its selection rate. A truly random procedure, which would have to draw its own selection rate in some random fashion, would be vastly inferior to the first assessor. In other words, the assessor is credited for choosing the right selection rate apart from relative validity. Similar arguments apply if the scale is continuous rather than dichotomous.

In conclusion, we propose to adopt L_{XY} as the most elementary and appropriate measure of likeness for bipolar proportion scales. The identity coefficient for absolute scales caters to the situation in which variables, for example, tests with different numbers of items, have arbitrary sizes and would be prevented from being perfectly associated because of that. No such provision is necessary with proportions. On the contrary, correction for size appears to detract from the representing of likeness between variables scored on a bipolar proportion scale.

Multivariate Structure

In this context, multivariate analysis is primarily a procedure for summarizing assessment data. On one hand, it is considered good practice not to ask assessors questions at a high level of abstraction but to spell out concepts by means of a questionnaire containing more concrete instances. On the other, the audience would not be served by receiving an answer sheet; the data had better be summarized. The obvious way of summarizing is to take an average (taking signs into account) of the scores on the relevant items. An objection, however, is that some items are more relevant than others so that weighting them would be more appropriate. The objection is of limited practical importance if the number of items is large, as weighted and unweighted sums are very much alike in that case. On the other hand, any gains from weighting are obtained for free with computerized scoring. Secondarily, multivariate analyses have theoretical spin-off: They provide insight into the multivariate structure of personality.

In the absence of external criteria for the differential relevance of questionnaire items, one would weight them according to their likeness to the (unweighted) total score on the relevant items. To keep weighted averages on the same bipolar proportion scale as the item scores, the weights w_j should be divided by the sum of their absolute values $\Sigma|w_j|$. On having assigned weights, however, the total score has been implicitly transmuted into a weighted average, so logic dictates weighting of the items according to their likeness to that weighted average. Consequently, an iterative process should be carried out, which ends on sufficient convergence. In practice, that procedure amounts to calculating the scores on the first principal component of the raw scores data matrix (see Horst, 1965). It maximizes the sum of the squared likenesses of the items with the weighted average. The second principal component does that with respect to the residual scores matrix, and so on. We thus conceive of principal components as weighted averages of variables and of principal component analysis as the way to find optimal weights. Finally, item loadings are calculated as likeness coefficients L_{XF} between item X and principal component F. MATLAB® (MATLAB Inc., 1994) routines for carrying out these procedures are available from Jos M. F. Ten Berge. Using standard programs, one would (a) transform scores onto the bipolar proportion scale, (b) calculate the matrix of average cross-products L_{XY}, (c) find the principal components of L, (d) find the corresponding component weights, (e) divide the weights for each component by the sum of their absolute values, (f) find the component scores, and (g) find the loadings L_{XF} on the principal components.

As an aside, we note that our approach resolves the dispute between proponents of a person centered (Magnusson, 1992) or typological conception of personality structure and the dominant variable-centered conception. Raw scores principal component analysis not only produces both matrices of factor scores and factor loadings, but contrary to standard

factor or component analysis, it is impartial, as it does not standardize scores in one direction (Q-analysis) or another (R-analysis). The bipolar proportion scale meets a classical (Cattell, 1944) plea for an interactive rather than either a normative or an ipsative conception of scores. As a matter of fact, one may conceive of the entries in a matrix of assessment data as likeness coefficients between individuals and variables, decomposable through principal component analysis. This rectangular or off-diagonal matrix of likeness or proximity relations (see Coombs, 1964) has the same principal component scores and loadings as the triangular matrices of likenesses between persons and likenesses between variables, respectively.

Another ramification is $N = 1$ principal component analysis. This novelty may find its way in situations in which the data do not form a matrix as in the common case in which each individual has been rated by a different set of assessors. In the minimal and most radical application, one individual has been rated by two assessors on one variable, say friendliness; the question is how these ratings are weighted in an optimal fashion. Take the following example:

	Assessor 1	Assessor 2	Average
Scores	1.0	−.5	.25

The first approximation to the average score is the unweighted average. To find the next approximation, we find the likenesses L between assessors and average:

Likeness	.25	−.125

We rescale the likenesses into weights with $\Sigma|w| = 1$:

Weights	.667	−.333,

so that the weighted average is now .833.

We could iterate the procedure, but in this elementary case, that would give the same weights, so convergence is immediate. What the example demonstrates, apart from the fundamentals of principal component analysis, is that the weighted average is drawn toward the more extreme assessment. As the likeness coefficient between a rating and the average rating is proportional to the extremeness of that rating, so is the weight. Note also that the second assessor receives a negative weight.

Evidently, analyses with small numbers capitalize heavily on chance. In the following, we present a less radical application in which the number of items is large.

REPRESENTING PERSONALITY

We applied the methodology set out previously to data collected by Hendriks (1997, p. 35, and following) in the process of constructing the Five-Factor Personality Inventory (FFPI; see also Hendriks, Hofstee, & De Raad, 2002). We used the ratings on a 5-point scale of 133 target persons, mostly first-year students of psychology at the University of Groningen, on 914 sentence items (e.g., "Keeps apart from others") by self and four others, forming 133 matrices of 5 × 914 assessments.

Weighting Assessors

All ratings were linearly transformed onto the bipolar proportion scale following Equation 1. For each of the 133 target persons, a 914 × 5 matrix of such scores was available. A raw scores principal component analysis using L as an index of association was applied to each of these 133 matrices to find optimal weights for averaging the five raters in question. The five assessors' weights on the first principal component were divided by the sum of their absolute values. The end result of this operation was a matrix of 133 × 914 averaged assessments of the target participants on the bipolar proportional scale.

The obtained weights are interesting by themselves. In the first place, only 2 of the 133 × 5 assessor's weights were negative (−.01 and −.02), which affirms the high quality of Hendriks' (1997) data. In the second place, self-ratings contributed slightly but significantly less than others' ratings: Whereas the overall average (absolute) weight is .20 as a consequence of the procedure, 84 self-weights were below that figure and only 49 above, $\chi^2(1, N = 133) = 9.21, p < .01$. If one accepts a definition of personality (cf. Hofstee, 1994) as the common component in the assessments of the population of relevant judges, these results point to a relative inferiority of self-reports amidst others' assessments. The result should be surprising to those who conceive of self as sharing more relevant information with others than do others amongst each other: Apparently, that pivotal position does not help enough. Theoretically, self-assessments could still have superior external validity; however, real-life criteria also tend to be in the hands of third persons.

Structuring Personality Assessments

The resulting 133 × 914 matrix of target individuals by items was in its turn subjected to raw scores principal component analysis using L as an association index. Five principal components were extracted to facilitate comparisons with the original solution. The sum of the absolute values of the 914 item weights per principal component was set to 1 to obtain component scores on the bipolar proportion scale.

In a standard principal component analysis using z scores and correlations, the first 5 eigenvalues were 210.2, 115.6, 56.6, 44.1, and 31.0. Thus, the first eigenvalue was 1.8 times as high as the second. In the raw-scores analysis, the first 5 eigenvalues of the matrix of L coefficients were 60.56, 9.62, 4.45, 3.40, and 2.56. Here, the first eigenvalue was 6.3 times the second. An additional size index in raw scores principal component analysis is $L_{FF} = \Sigma F^2/N$, the mean of the squared component scores. For the first 5 principal components, the sizes are .0866, .0156, .0081, .0056, and .0045. According to

this index, the first principal component is 5.6 times as big as the second. We emphasize that these high figures are not an artifact in the sense of an automatic consequence of adopting a particular scale. That would be the case with a unipolar scale leading to all-positive L_{XY}s. With a bipolar scale, however, there is no such restraint: If item scores would be symmetrically distributed around the zero point, results from raw scores principal component analysis would not differ much from standard outcomes. The dramatic increase in the relative contribution of the first principal component reflects the dominant role of individuals' social desirabilities in personality assessments, which is partly suppressed in standard analysis.

On interpreting these scores, their size should be kept in mind. Scores between –.25 and +.25 are in the neutral range indicating neither a positive nor a negative likeness between the individual and the principal component. Of the 133 participants, 47 (35%) had neutral scores on all five principal components. Of the remaining scores, none even approached the extreme (> +.75 or < –.75) ranges; only two scores were (slightly) above .5.

Of the 86 participants with nonneutral scores, 79 (92%) had their highest absolute score on the first principal component, which may be interpreted as (un)desirable social behavior; it is mostly negatively defined by items such as takes advantage of others, abuses people's confidence, insults/hurts people, treats people as inferiors, and cuts others to pieces. Most of these items were markers for the negative pole of Factor 2, Mildness, in Hendriks' (1997) study. Note that this is not social (un)desirability in the self-presentational sense but undesirable social behavior in a moral perspective. Of the 79 nonneutral scores on this component, the large majority (75, being 95%) were positive; 4 were negative, ranging from –.27 to –.45. Of all 133 participants, 127 (again 95%) had positive scores on the first principal component. If this sample is at all representative, our Orwellian saying about people's desirabilities has to be qualified somewhat: By far the most people are weakly to mildly socially desirable; a few are weakly to mildly undesirable.

Four participants had nonneutral scores on the second principal component of which three were positive and one negative. A number of 56 (out of 914) items had their highest loading on this component, constituting a self-willed type that takes charge, wants to pull the strings, makes his/her own rules, wants to have it his/her own way, seeks confrontations, and knows how to manipulate a situation. Three participants had nonneutral scores on the third principal component of which two were negative. On this component, only 17 items had their highest loading, such as needlessly worries a lot and is afraid that he/she will do the wrong thing, versus readily overcomes setbacks and can take his/her mind off his/her problems; according to Hendriks' (1997) research, the latter two are marker items for Emotional Stability. On the fourth and fifth principal component, all 133 scores were in the neutral range.

As might be expected given the large size of the first principal component, varimax rotation of the matrix of component scores did not bring about great changes; all diagonal elements in the rotation matrix were above .8. The number of participants with neutral scores on all components rose from 47 to 54. Of the 70 participants with a nonneutral score on the first rotated component, 4 were negative (these belong to the same participants as before rotation). The numbers of participants having their highest absolute score on the second through fifth rotated component were 3, 3, 2, and 1, bringing the total back to 133. Before and after rotation, 3 participants had nonneutral scores on both the first and the second component, so the rotation did also not result in a simpler person structure.

Single-Source Analysis: Self-Assessments

We consider aggregation over assessors to be the standard for personality assessments, if only to compensate for assessor error. However, single-source perceptions of personality may be of interest as such, for example, if self-assessments are to be confronted with other assessors' points of view as in therapeutic or personnel management settings. Evidently, the findings from the aggregated data cannot be automatically generalized to single-source data. We therefore carried out a supplementary analysis on just the 133 self-ratings in the Hendriks (1997) study using the same methodology.

Using standard principal component analysis, the first 5 eigenvalues were 112.3, 84.6, 48.6, 42.4, and 32.0. Therefore, the first eigenvalue was 1.33 times the second. In the raw-scores analysis, the first 5 eigenvalues were 47.47, 14.42, 7.78, 6.50, and 5.51. Therefore, the first eigenvalue was 3.33 times the second. The sizes $L_{FF} = \Sigma F^2/N$ of the first 5 principal components were .0688, .0223, .0135, .0109, and .0095. Thus, according to both criteria, the first principal component was more than three times as large as the second. The relative increase of the first principal component as a result of absolute scaling is less than in the aggregated data but still sizeable.

Of the 133 participants, 71 (53%) had neutral scores on all five principal components. This is far more than the 35% in the aggregated data and reflects the smaller sizes of the principal components in the self-assessments. Still, the component scores ran higher, the three largest scores being between .65 and .70. Thus, any extreme item scores that were present in the self-assessments were not smoothed out as much as happens on aggregation over assessors; but even in these data, extreme (> +.75 or < –.75) component scores did not occur.

Of the 62 participants with nonneutral scores, 44 (71%) had their highest absolute score on the first principal component; for the other components, the numbers were 10, 3, 2, and 2, respectively. The interpretation of the principal components in terms of their highest loading items remained virtually unchanged. Clearly, however, the dominance of the first principal component was somewhat mitigated in the

self-assessments. Of all 133 scores on the first principal component, only 4 (3%) were negative; none of the nonneutral scores were. These figures were in line with the results from the aggregated analysis.

The single-source analysis was based on only 133 assessors, so not all aspects of the results should be expected to generalize to other samples. The main results from the aggregate analysis, namely, absence of principal component scores in the extreme regions and a large relative boost of the first principal component, are replicated in the self-assessments, although the effects are less extreme than in the aggregate analysis.

DISCUSSION

Our procedures render an unromantic turn to the study of personality. Little remains of the shades and subtleties of individual differences in temperament and character. By far the most people appear to be faintly to mildly okay, a few are not, and a handful (in the self-assessments, a sizeable minority) may better be characterized in other terms; that is about all there is to it. It is as if we are reminded of the fact that we share 99% of our genes with primates and almost all with one another. How do these outcomes relate to our educated intuitions about personality?

Size of Individual Differences

Perhaps the most counterintuitive outcome is that extreme scores were not observed: Even on varimax rotation of the component scores matrix, the highest score of all fell short of .55 in the aggregate analysis. Few persons should thus be expected to be more than halfway agreeable, conscientious, and the like. However, don't we all know individuals who seem to be much more extreme than that in one respect or another? The answer is undoubtedly affirmative, for if assessors rate individuals directly on traits, a sizeable fraction of these ratings will be at the most extreme scale points.

However, the contradiction is easily resolved. First, assessors do not agree all that much. On averaging their ratings, regression toward the midpoint of the scale is virtually automatic. It occurs even if extreme assessments receive greater weights as was the case in this analysis. Who is right, the individual assessor or the collective? That depends on the frame of reference. In a poetic or romantic context, truth is subjective; one's individual judgment is all that counts. However, in an intersubjective context, we have to account for the well-established fact that even our own judgments about ourselves and about others had better be subjected to statistical regression. Hendriks' (1997) exemplary data involving five assessors per target are much more relevant and authoritative in this respect than our intuitions, scientifically speaking.

Second, the component scores in this study came about as a weighted average over 914 items or facets of personality-relevant behavior, each of which triggers another aggregation operation at a more specific level. Human behavior is not very consistent. In the supreme autonomy of our private intuitions, we can easily discount the fact that the prototypical extravert is seen sitting in a corner. However, scientifically speaking, we can only acknowledge that a score of .5 corresponding with a 75–25 preponderance of positive over negative instances of a trait is probably a whole lot. On second thought and in view of the additional imperfect assessors' reliability, one may even become distrustful the other way and wonder whether such high scores do not represent capitalizations on chance (which they probably do to some extent).

Personality Structure

Next, the results run counter to the bigness of factors beyond the first principal component. The dominant impression of a highly differentiated, five-dimensional structure (see, e.g., De Raad, 2000; De Raad & Perugini, 2002) came about through the combined use of standard scores and varimax rotation, which spreads their variance over factors. With the bipolar proportion scale, the results are varimax resistant. That is not because the content of the principal components changes much on using an absolute scale. We calculated a standard Big-Five solution (five principal components plus varimax) on Hendriks's (1997) data and carried out a multiple regression for each of these factors using the five raw scores principal components as predictors. The multiple correlations ranged between .983 and .999, indicating excellent coverage of the standard Big-Five space.

This change in perspective is reminiscent of what happened to the concept of intelligence. In the tradition of Thurstone (1938) to Guilford (1967), intelligence was conceived as multifaceted and complex. Toward the end of the 20th century, the hierarchical conception with *g* at the top took over (see, e.g., Herrnstein & Murray, 1994). The intercorrelations of personality items, if scored in socially desirable direction, will be found to be in the same order of magnitude as the intercorrelations of intelligence items. Thus, even on adhering to relative scales and corresponding statistics, a general *p* factor of personality (Hofstee, 2003) seems to deserve serious consideration. As with intelligence, the implication is not that there is nothing more to personality than the first principal component. The implication is that other dimensions are of secondary importance. The overall picture that arises is a positive manifold, bipolar in the case of personality assessment, according to which lower level concepts form an oblique structure: a sort of multidimensional double cone (Peabody & Goldberg, 1989) but with a spatial angle of less than 90°.

As we noted at the beginning of this article, relative scales and statistics are relevant in comparative contexts; therefore, so are well established trait structures such as the Big Five. It thus looks as if we have two conceptions of trait structure,

one relative and the other absolute. Could the true structure lie in between? That question amounts to asking for contexts that have both relative and absolute features. Those situations may well constitute the general case. Even in comparative personnel selection, one would like to make sure that the best candidate is at all fit for the job; in selecting a prizewinner, the jury may decide not to award the prize. On the other hand, the central tendency of a trait's distribution tends to induce anchoring effects and therefore, a relative component. So both strictly absolute and relative scaling may be found to be extreme cases. The development of mixtures of absolute and relative scales is beyond the scope of this article. At this stage, it may suffice to suggest that standard structures based exclusively on relative scaling are not located in the middle of the road but at its edge, the other edge being occupied by the bipolar proportion model.

ACKNOWLEDGMENTS

The content of this article was presented as an invited paper in the symposium on Personality Types Versus Personality Dimensions at the Eleventh European Conference on Personality, July 2002, in Jena, Germany. We have profited a great deal from incisive comments by anonymous reviewers.

REFERENCES

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506–520.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90,* 105–126.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51,* 292–303.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

De Raad, B. (2000). *The Big Five personality factors: The psychological approach to personality.* Göttingen, Germany: Hogrefe.

De Raad, B., & Perugini, M. (2002). *Big Five assessment.* Göttingen, Germany: Hogrefe.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Hendriks, A. A. J. (1997). *The construction of the Five-Factor Personality Inventory (FFPI).* Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands.

Hendriks, A. A. J., Hofstee, W. K. B., & De Raad, B. (2002). The Five-Factor Personality Inventory: Assessing the Big Five by means of brief and concrete statements. In B. de Raad & M. Perugini (Eds.), *Big Five assessment* (pp. 79–108). Göttingen, Germany: Hogrefe.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve.* New York: Simon.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8,* 149–162.

Hofstee, W. K. B. (2002). Types *and* variables: Towards a congenial procedure for handling personality data. *European Journal of Personality, 16,* 89–96.

Hofstee, W. K. B. (2003). Structures of personality traits. In I. B. Weiner (Series Ed.) & T. Millon & M. Lerner (Vol. Eds.), *Handbook of psychology: Vol. 5. Personality and social psychology* (pp. 231–254). Hoboken, NJ: Wiley.

Hofstee, W. K. B., & Hendriks, A. A. J. (1998). The use of scores anchored at the scale midpoint in reporting people's traits. *European Journal of Personality, 12,* 219–228.

Hofstee, W. K. B., Ten Berge, J. M. F., & Hendriks, A. A. J. (1998). How to score questionnaires. *Personality and Individual Differences, 25,* 897–909.

Holley, J. W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24,* 749–753.

Horst, P. (1965). *Factor analysis of data matrices.* New York: Holt.

Magnusson, D. (1992). Back to the phenomena: Theory, methods, and statistics in psychological research. *European Journal of Personality, 6,* 1–14.

MATLAB, Inc.. (1994). MATLAB (Computer software). Natick, MA: Author.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Peabody, D., & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology, 57,* 552–567.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1.* Chicago: University of Chicago Press.

Zegers, F. E., & Ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika, 50,* 17–24.

Willem K. B. Hofstee
The Heymans Institute
University of Gronigan
Grote Kruisstraat 2–I
9712 TS Groningen
The Netherlands
E-mail: w.k.b.hofstee@ppsw.rug.nl

Received November 4, 2002
Revised October 2, 2003

Theory Into Practice, 48:63–71, 2009

ISSN: 0040-5841 print/1543-0421 online
DOI: 10.1080/00405840802577627

Routledge
Taylor & Francis Group

Gregory J. Cizek

Reliability and Validity of Information About Student Achievement: Comparing Large-Scale and Classroom Testing Contexts

Reliability and validity are two characteristics that must be considered whenever information about student achievement is collected. However, those characteristics—and the methods for evaluating them—differ in large-scale testing and classroom testing contexts. This article presents the distinctions between reliability and validity in the two contexts and provides recommendations for enhancing the quality of information about student achievement, with particular attention to classroom decision making.

Gregory J. Cizek is a professor of Educational Measurement and Evaluation at the University of North Carolina at Chapel Hill.

Correspondence should be sent to Gregory J. Cizek, Educational Measurement and Evaluation, School of Education, CB 3500, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3500. E-mail: cizek@unc.edu

CONCERN FOR RELIABILITY and validity applies whenever information is gathered. In classrooms, that information is often a test score, but it is too narrow to consider reliability and validity only in that sense. This article examines reliability and validity in two contexts: large-scale testing and classroom testing situations where the focus is student achievement in elementary and secondary schools.

What Is a Test?

Test is a widely misunderstood concept. Commonly—but too narrowly—test connotes a collection of multiple-choice questions, bubble sheets, and rigid administration conditions. Although multiple-choice questions administered in a standardized way and scored by optical

scanners qualify as a test, it is only one of many possibilities. In fact, for many teachers, this configuration may rarely be used.

Broadly conceived, a test is any systematic sample of a person's knowledge, skill, or ability. Thus, many things qualify as a test. However, specific steps must be taken so that the sampling yields dependable and accurate information. In large-scale contexts, for example, prescriptive administration conditions facilitate trustworthiness of the information, and permit comparisons of information gathered across different persons, settings, or occasions. In classroom situations, the specification of conditions changes because the classroom context is inextricably linked with the information-gathering procedure itself (Brookhart, 2003). However, that linkage does not mean that context and assessment are hopelessly confounded. Rather, it is desirable that even the most informal classroom assessment yields information about student performance that is generalizable to other contexts. Otherwise, would any real learning have occurred?

A second aspect of a *test* is that it is only a sample of what a test-taker knows or can do. It is impossible or impractical to observe everything. Tests capture only a small portion of what could be observed, and the interpretation of test results requires making an inference from the observed performance to the larger domain of interest. For example, it is not usually of interest that a student can correctly respond to the question: "4 × 8 = ?" Instead, a teacher would like to extrapolate from a student's correct answers to 10 such questions that the student has a good grasp of the basic multiplication facts. The act of going from observed test results to a conclusion about a student's level of knowledge or skill is called *inference*.

There are two conclusions to be drawn regarding *test* and *inference*. First, the meaning of *test* is broad and essentially unrelated to any specific format. Thus, each of the following could be a test: an observation of student cooperative group work habits; a checklist of proper technique for working with clay, a survey of students regarding the extent of bullying on the playground; oral probing of the steps a student took to solve a mathematics problem. In short, as long as something is a systematic sample of behavior, it could be a test. Second, interpretation of test results requires *inference*. The more comprehensive the sample of student work and the more carefully conducted the observation, the more confidence we can have that the inference is plausible.

There are two corollaries to these conclusions. First, because they rely on a sample of information, all inferences are tentative. Although some inferences can be made with more confidence than others, no conclusion about what a student knows or can do is ever made with certainty. Regardless of how carefully the information is gathered, there is always the possibility that conclusions will be incorrect. Second, although the terms *test* and *assessment* are often used as if they were synonymous, they are not. Whereas a test is a single, systematic collection of information, an assessment is much broader. Assessment refers to "the planned process of gathering and synthesizing information relevant to the purposes of: discovering and documenting students' strengths and weaknesses; planning and enhancing instruction; or evaluating and making decisions about students" (Cizek, 1997, p. 10).

An analogy may help clarify the distinction. A person who comes to an emergency room for treatment might have blood drawn (a test), blood pressure taken (another test), and so on. All of the test results are then scrutinized in order to come to a conclusion about possible diseases or conditions, and to develop potential courses of treatment (an assessment). Likewise, educational assessments are based on multiple, often diverse sources of information—much of which is typically gathered using tests. A good example of an assessment is the process engaged in by an Individualized Educational Program (IEP) team.

Purposes of Testing

When examining any testing context, the foremost consideration is the purpose of the test. A test developer may be very clear about the purpose of a test, or the purpose may be

overstated, or understated, or not articulated at all; regardless, an intended purpose exists. The purposes of large-scale and classroom tests are usually different.

Large-scale student achievement tests like those required by the *No Child Left Behind* Act (2001) typically have two purposes: (a) global estimations of an individual student's competence in an area; and (b) accountability for student progress at an aggregated level, such as classrooms or districts. The first (individual) purpose can be seen in test results in which categorical classifications for individual students are reported (e.g., Basic, Proficient, Advanced). The second (system monitoring) purpose is evident in reports related to Adequate Yearly Progress. To accomplish their purposes, these tests are developed, administered, and scored according to strict standards that routinely yield high quality information.

Classroom tests usually have different purposes than large-scale tests. Classroom testing is a broad concept, capturing such diverse information gathering systems as observations, preconstructed tests in teacher's guides, informal teacher–student interactions, teacher-constructed tests, and others. Like any test, classroom testing must be considered in terms of an intended purpose. Some tests serve primarily to support grading decisions. Tests can be used to motivate students or to promote studying. Other tests may be used primarily to provide formative feedback, or as diagnostic measures for determining specific strengths and weaknesses of individual students. Still others may be intended primarily as an aid to the teacher in instructional planning, grouping, or pacing of learning activities.

Classroom tests have the potential to accomplish these purposes very well and to provide accurate and useful information. However, like their large-scale counterparts, they must be designed with a specific purpose in mind and must be carefully constructed, administered, and scored. Both groups—developers of large-scale tests and educators who develop classroom tests—must begin the process with a clear statement of the intended purpose(s) that a test aspires to accomplish; that statement is essential for gauging the extent to which the test is adequate.

In summary, a test is any systematic sample of information, and includes diverse formats, procedures, and information-gathering protocols. Test results are necessarily tentative, because inference is always required. An assessment is more comprehensive than a test. All tests are designed with a specific purpose in mind. The intended purpose of a test should be explicitly stated and is the starting point for considering the technical adequacy of any test. With these key ideas in mind, we now turn to how the quality of large-scale and classroom tests is evaluated.

General Principles for Evaluating Test Quality

Information-gathering procedures (i.e., tests) are evaluated according to two criteria: reliability and validity. Conceptually, the two criteria apply with equal force and importance to both large-scale and classroom tests; however, the procedures used to establish evidence of reliability and validity differ across those settings.

Reliability

Reliability refers to the dependability of test results. An information-gathering procedure should yield results that, when repeated under similar conditions, should yield the same or highly similar results. Although reliability is not considered to be the most important aspect of test results—this honor is reserved for validity—it is an essential first step in ensuring that test results have credibility and usefulness. If test results aren't dependable, then it doesn't even make sense to speak about their validity.

An everyday example can help to illustrate reliability. Suppose a person weighed himself at home on a typical scale. One morning, the person obtained three consecutive measurements: 280, 57, 183 pounds. Startled, the person repeated the procedure several more times, each time obtaining dramatically different readings. What can be said about this situation? Assuming that

the conditions of the measurement were stable (e.g., the person wasn't holding barbells on one occasion and not another), we might conclude that the scale was worthless. Such fluctuations would yield no dependable information about the person's weight. The best course of action would be to discard the scale and purchase a new one.

A desirable scale is one that, at minimum, gives consistent results on similar occasions. Suppose that the same person purchased a high-quality scale and again performed a series of consecutive measurements, yielding readings of 183, 185, and 184 on consecutive measurements. Small differences like these would be expected, even when a very good measuring device is used. The same principle is true for educational measurements. Because all educational tests consist of a sample of observations and because both the students who respond to the test and those who score the responses are susceptible to various unpredictabilities (which psychometricians call *random errors*), no score can be considered to be a perfectly dependable snapshot of a student's performance. One of the principal aims of testing specialists is to eliminate, to the extent possible, even small differences in results attributable to these random errors.

In large-scale contexts, the focus of reliability is on the dependability of test results, although the procedures for estimating that dependability depend on the kind of unpredictabilities that are of interest. Each procedure invokes a statistical procedure that results in a *reliability coefficient* that can range from zero (0.0) to one (1.0). A reliability coefficient of 0.0 would signify scores that are completely undependable (like the first scale described previously!); a reliability coefficient of 1.0 would indicate perfect dependability and the potential for the test scores to be used with great confidence. Coefficients between 0.0 and 1.0 indicate relatively poorer (nearer to 0.0) or greater (nearer to 1.0) dependability. Realistically, reliability coefficients of 0.0 or 1.0 are never observed, although the reliability values for high-quality tests (such as statewide assessments, the *SAT*, the *Iowa Tests of Basic Skills*, etc.) are routinely found to be .90 or higher, signaling that scores on those tests are highly dependable.

As previously indicated, methods for estimating dependability in large-scale contexts depend on the likely source of score fluctuations that are of greatest concern. For example, if two equivalent forms of a test were developed, knowing to what extent scores on the two forms could be used interchangeably would be of interest. That is, do the different forms yield consistent results? The reliability estimate might reveal that the two forms could be considered essentially identical (reliability near 1.0) or that choosing one form or the other could make a difference in a student's score (reliability near 0.0). This kind of reliability estimate is called a *coefficient of equivalence.*

Alternatively, it might be of interest whether scores on a single test would likely be different if the same test were repeated on different occasions. That is, are the test results dependable over a certain time interval? The appropriate reliability estimate in this situation might reveal that performance on the test is highly stable over time (reliability near 1.0) or that the intervening time has a large influence on students' scores (reliability near 0.0). This kind of reliability estimate is called a *coefficient of stability* or *test–retest* reliability.

One commonality in the two previously described situations is that students must take two tests—either the same test on different occasions or two equivalent test forms—to obtain a reliability estimate. An alternative is an approach called *internal consistency*, which includes Cronbach's alpha, KR-20, and other methods. These procedures attack the question of dependability by analyzing the extent to which items within a test are related. The stronger the relationships among items (i.e., near 1.0), the more dependable the results; the weaker the relationships (i.e., near 0.0), the less dependable the results.

The concept of reliability is the same, but different, in classroom assessment contexts. It is the same in that the test, observation, or sampling of student performance should yield information that is dependable. It is the manner in which that information is gathered and summarized that differs.

Classroom situations are not usually amenable to large-scale reliability estimation procedures.

First, many reliability estimation procedures developed for large-scale context are inappropriate when students do not respond to items or tasks independently, where observation purposes or contexts vary, or where the tasks comprising an assessment are not independent. In classroom contexts, student performances or observations may have been purposefully designed to be gathered in situations where the students work with teachers or other students. Second, the sample sizes in classrooms are usually small enough that, for statistical reasons, any resulting reliability coefficients would be unreliable themselves! Unfortunately, although testing specialists have refined the procedures for estimating reliability in large-scale contexts, much less attention has been devoted to methods of gauging reliability in classroom contexts.

Although many methods of estimating reliability are inappropriate for classroom settings, the same concerns are still of interest. For example, the reliability information provided by a test–retest procedure answers the question: "Would the student's results likely differ if he or she was observed performing the task again?" In classroom contexts, an identical test would not be administered twice to the same students to gather reliability information. However, a variation of the test–retest procedure is done by teachers. Teachers know many factors can cause students' performances on any given occasion to be better or worse than what they are capable of doing. These unpredictabilities are the reason that few language arts teachers would assign a grade in a Speech class based on only one attempt, why few science teachers would make conclusions about students' knowledge of chemistry based on a single lab report, and so on. Even in one-on-one conferences with students, a teacher interested in dependable information does not settle for a student's answer to a single question; the teacher probes, or asks a follow-up question to ascertain the dependability of the information gleaned from the student's first response. It is not necessary to quantify this in a statistical sense, and such methods can legitimately be seen as a kind of test–retest procedure with the clear aim of helping the teacher understand the dependability of any conclusions about the student.

Another kind of reliability that applies to classroom contexts is *scoring reliability.* Scoring reliability is relevant when teachers rate students' speeches, grade term projects, score students' responses to essay questions, and so forth. In these cases, the student work might vary substantially in features (e.g., legibility, organization, tardiness) that are unrelated to the primary characteristic the teacher intends to evaluate. In addition, the teacher also becomes a source of unpredictability, due to variation in time pressures, mood, leniency/stringency, and other factors that affect scoring.

There are two kinds of scoring reliability: *intrarater consistency* and *interrater agreement.* Intrarater consistency is the dependability with which a teacher scores students' work across the sample of papers, performances, essays, and so on. It is consistency within a scorer. Ideally, a teacher would apply a rubric or scoring procedure in a consistent fashion, unaffected by students' handwriting, fatigue, and so on. There are several strategies for bolstering intrarater consistency. For one, many teachers insist that all assignments be turned in at the same time and they set aside a block of time to apply fixed criteria and a uniform perspective to the scoring process. For another, many teachers prescribe certain requirements for all work submitted (e.g., typed, double-spaced) to help rule out potential sources of irrelevant variation in scoring. Finally, teachers can set aside time to review scores after making initial judgments. For example, with prior grades masked, a sample of papers can be rescored to determine if grading standards have drifted.

Interrater agreement is the extent to which different raters of the same student work agree on the score for the work sample. It is consistency across scorers, and involves having two or more persons rate the same student work. In many classroom contexts, investigating interrater agreement is unnecessary (e.g., scoring a weekly spelling test), but when an assignment carries important consequences, it would be desirable to assess interrater agreement. A teacher could do this by, for example, asking a colleague to apply

the same rubric to a subset of student essays and comparing the scores. Disagreements regarding the scores would indicate unreliability in scoring, uneven application of the scoring guidelines, or systematic leniency or stringency in grading.

Validity

Validity is considered "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9). Validity refers to the degree to which the inferences yielded by a sample of behavior (e.g., assignment, quiz, observation, interview, etc.) are accurate. Less technically, validity is the degree to which judgments about a student's knowledge or skill are supported by adequate evidence. It is the degree to which conclusions based on test information are on target.

The ways in which validity evidence is gathered are fairly well-developed in large-scale testing contexts. Large-scale validation often involves statistical techniques that are not well-suited to classroom contexts. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) describe validity evidence based on: (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. Rather than describe these categories in detail, it is sufficient to note that large-scale validation typically taps multiple sources and summarizes the validity evidence into an overall judgment about the degree to which intended inferences are supported. For example, a state test purporting to address higher-order thinking skills in fourth-grade mathematics would likely demonstrate alignment between test items and the state curriculum (evidence based on test content); think-aloud studies might support that the test actually required higher-order thinking and not merely recall or calculation skill (evidence based on response processes); and statistical analyses might reveal that performance on the test was strongly related to performance on other mathematics tests and weakly to Reading and Writing test results (relations to other variables). It may now be obvious that validation for large-scale tests is a substantial endeavor.

Classroom testing situations are not usually amenable to many of the large-scale validation procedures because classroom conditions are often variable across students, and students may perform different tasks, answer different questions, be observed in different contexts, receive differing amounts of prompting, assistance, or collaboration, and so on. Typical validation for classroom tests may be less formal than large-scale contexts, but for classroom assessment procedures to produce high-quality information, it is still essential that validity information be gathered and evaluated.

Given the ubiquity of classroom assessment, one might assume that validity procedures for classroom assessments are well-developed. Regrettably, great progress has not been made in formalizing the sources of validity evidence that are appropriate to classroom contexts, nor have key competencies in classroom assessment validation been integrated into educator preparation programs. Nonetheless, it is still possible to suggest some of the important sources of validity information that are available to teachers and that should be collected and documented to support important decisions about students.

Among the important validity concerns that apply to classroom testing are: content validity, bias, appropriate accommodations, alignment of the assessment to relevant content standards or curricular goals, adequate opportunity for students to learn the tested material, and the sensitivity of test results to instructional interventions. Most of these validity concerns also apply to large-scale tests, but they are addressed in distinctly different ways in the classroom context. A summary of sources of validity evidence for classroom tests is provided in Table 1.

A few of the sources of validity evidence shown in the table should be highlighted. Consider the source called *Alignment*. One aspect of classroom test validity is the extent to which a test mirrors the instructional goals a teacher has emphasized. After constructing a quiz, assignment, or other test, teachers can enhance validity by explicitly comparing the teacher's

Table 1
Sources of Validity Evidence for Classroom Tests

Source of Evidence	Validity Question(s)	Description
Content	What content domain is covered by the test?	Involves deciding what test results should represent (the intended inference) and communicating to students what a test will cover.
Alignment	Is the test content drawn exclusively from the intended domain? Do the test items or tasks match the content standards?	Involves ensuring that the test is completely matched to the intended domain (all aspects of the intended domain are covered; no areas are covered too much or too little).
Item/task construction	Are the test questions or tasks that students are asked to perform clear? Are directions clear?	Involves review of all test questions, prompts, directions, and other student materials for appropriate level, clarity, correct grammar, spelling, and adherence to accepted test development guidelines.
Administration conditions	Are the conditions under which students are tested conducive to them performing their best? Are time constraints appropriate?	Involves ensuring that testing conditions are comfortable, free of distractions, and that timing does not inappropriately affect students' scores (unless the intended inference is students' ability to perform under time pressure).
Scoring	Was the test scored appropriately?	Involves development of scoring rubrics, answer keys, observation protocols, etc., that are accurate, aligned with test directions, and applied consistently and objectively.
Opportunity to learn	Have all students had sufficient opportunities to learn the knowledge or skills being tested?	Involves ensuring that the tested material has been covered in classroom instruction, activities, homework assignments, lab experiences, or other contexts, so that all students have had adequate time and opportunity to master the content that is tested.
Fairness	Are the test questions or tasks fair for all students? Are there any aspects of the test or procedures that would unfairly advantage or disadvantage some students?	Involves teacher review of items/tasks/ assignments, etc. to ensure sensitivity to differences in students' gender, culture, ethnicity, language diversity, etc., and that any potentially insensitive language, bias, or sources of unintended score differences are eliminated.
Accommodations	Are assessment accommodations for students with special needs in place? Do accommodations foster students' demonstrations of their real levels of knowledge or skill? Do any accommodations provide an inappropriate advantage to students?	Involves ascertaining what accommodations are necessary for students to perform according to their true levels of knowledge or skill and implementing these accommodations during testing. Also involves ensuring that any accommodations do not provide an unfair advantage or alter the intended inference.
Instructional sensitivity	Do test results primarily reflect influences of instructional experiences and student learning?	In pre- and posttest contexts, involves review of student performance to determine if improvements are due primarily to instruction, as opposed to development, maturation of students, or other irrelevant factors.

learning goals to the items or tasks to ensure that they reflect the instruction and learning activities. When misalignment occurs because a test does not have high fidelity to classroom instruction, the accuracy of inferences is degraded. Or, in the words of students, the test is "Unfair!"

The topic of alignment is important in both large-scale and classroom assessments. There are three major components to alignment: curriculum, instruction, and assessment. In the large-scale context, we expect that a statewide assessment will be crisply aligned with the state's prescribed curriculum or content standards. Most states work to ensure a tight linkage between curriculum and state tests, and many states provide professional development opportunities and resources for teachers to help ensure that the third, and essential component of the system—instruction—also aligns to the intended goals. In the classroom context, a teacher must work to ensure that his or her goals for instruction are explicitly linked in the same way: through explicit articulation of those goals in instructional planning, careful integration in instructional activities, and fidelity to the goals in any assessments of student learning.

Consider also the source of validity evidence called *Fairness* in Table 1. Fairness refers to the extent to which there is an absence of factors, unrelated to the intended purpose of a test, that advantage or disadvantage students. The presence of such factors is called bias. For example, poor readers would be disadvantaged on a reading test, and on a mathematics test consisting of complex story problems. The former situation would not constitute bias; the latter would. Likewise, students' performances on classroom measures should not depend on shared culture, common experiences, language, or other factors that are irrelevant to whatever the teacher intends to assess. Payne (2003) has recommended the simple strategy of asking a colleague to review a test before administering it to students; in addition to evaluating the match between intended goals and the test items/tasks (i.e., alignment), a colleague can often offer constructive comments on potential sources of bias.

Finally, consider the source of validity called *Scoring* in Table 1. Scoring validity refers to the accuracy of assigned scores. One common threat to score validity in classrooms is called the *halo effect*. Here is how it operates: Imagine that a teacher (using a rubric to promote scoring consistency) is scoring a test comprising five essay questions. The teacher's grading procedure involves reading all five of one student's essay responses, before proceeding to grade other students' responses. Now imagine that the student's response to the first essay was outstanding. A teacher might subconsciously be expecting a very good answer from that student to the second question, and so on. Without intending to do so, the teacher might assign a higher grade to the student's second response than the essay deserved. This phenomenon, the halo effect, also operates in the other direction: A poor first response sometimes creates an unintended expectation that the student's subsequent response will also be weak. To avoid this validity threat, teachers can simply rate all of the responses to question 1, then rate all of the responses to question 2, and so forth.

Conclusions

Classroom assessment has not received the same attention as large-scale testing. The prominence of large-scale testing has overshadowed the role and importance of classroom assessments, and there are logical reasons why this has occurred. The considerable financial resources allocated to large-scale testing warrants greater attention to ensure that the investment is well-placed. The consequences of large-scale testing are often greater than those associated with classroom assessments; thus, greater attention is paid to ensuring that large-scale tests accomplish their purpose as dependably and accurately as possible. And, by definition, large-scale assessments affect many people, and the data yielded by large-scale tests facilitates important public debates about educational progress and reform (Cizek, 2007).

However, it is likely that large-scale testing programs are reaching a point of diminishing returns as regards their potential to stimulate greater gains in student achievement. Research and development in the area of education measurement has begun to focus on the great promise of classroom assessment. For example, Black and Wiliam (1998) have highlighted the important role that classroom assessments serve in gathering information that teachers can use to adapt their teaching to the immediate needs of students. Their work has also demonstrated that well-designed and -implemented formative classroom assessments can have profound effects on student achievement, particularly for low-achieving students and students with special needs.

For classroom assessment practices to fulfill their potential, however, they must yield high-quality information that is stable (i.e., reliable), crisply focused on a purpose (i.e., valid), and meaningful in a broader context than the one in which the information was collected (i.e., generalizable). Focusing on these characteristics will help enhance the utility, importance, and power of classroom assessments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7–74.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5–12.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 1–32). San Diego, CA: Academic.

Cizek, G. J. (2007). Formative classroom assessment and large-scale assessment: Implications for future research and development. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 99–115). New York: Teachers College Press.

No Child Left Behind Act. (2001). P. L. 107-110, 20 U.S.C. 6301.

Payne, D. A. (2003). *Applied educational assessment, 2nd edition.* Belmont, CA: Wadsworth.

JIDR Journal of Intellectual Disability Research
Published on behalf of mencap and in association with IASSID

Journal of Intellectual Disability Research doi: 10.1111/j.1365-2788.2011.01396.x

VOLUME 55 PART 5 pp 453–461 MAY 2011

Role of assessment tests in the stability of intelligence scoring of pre-school children with uneven/delayed cognitive profile

P. Yang,[1] Y-J. Jong,[2,3,4] H-Y. Hsu[2,3] & F-W. Lung[5,6,7,8]

1 *Department of Psychiatry, Kaohsiung Medical University and Kaohsiung Medical University Hospital, Kaohsiung, Taiwan*
2 *Department of Pediatrics, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan*
3 *Department of Laboratory Medicine, Kaohsiung Medical University Hospital, Kaohsiung, Taiwan*
4 *Graduate Institute of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan*
5 *Department of Psychiatry, Kaohsiung Armed Forces General Hospital, Kaohsiung, Taiwan*
6 *Department of Psychiatry, National Defense Medical Centre, Taipei, Taiwan*
7 *Department of Neurology, Kaohsiung Medical University, Kaohsiung, Taiwan*
8 *Calo Psychiatric Centre, Pingtung County, Taiwan*

Abstract

Background As part of an ongoing clinical service programme for pre-school children with developmental delay in an Asian developing country, we analysed the effect of three assessment tests, that is, Bayley Scale of Infant Development-II, Leiter International Performance Scale – Revised and Wechsler Preschool and Primary Scale of Intelligence – Revised – Chinese, on the stability of intelligence quotient (IQ) of children from pre-school through early childhood.

Methods The participants were 313 Taiwanese pre-school children with uneven or delayed cognitive profile and they were followed through early childhood. IQ stability was explored by different tests and among children of different clinical diagnosis: 168 children with non-autistic intellectual disability, 73 children with autism spectrum disorder, 58 children with mixed receptive-expressive language disorder and 14 children of other heterogeneous diagnoses. Stability of scores was evaluated using the *r*-squared for Pearson's coefficients to see the correlation between initial IQ (IQ1) and follow-up IQ (IQ2). Multiple linear regressions were also applied to see whether IQ1 had predictive ability for IQ2 and test–test difference in the total 313 children and each diagnostic subgroup.

Results Results revealed that mean IQ1 was 65.8 ± 15.4 while mean IQ2 was 73.2 ± 17.9 for the total 313 children. The IQs were stable across an average follow-up duration of 38.6 ± 22.1 month from pre-school into early childhood. Patterns of positive correlations between IQ1 and IQ2 were noted by all the tests (*r*-squared = 0.43–0.5, all $P < 0.001$) and in the majority of diagnostic subgroups. Multiple regressions analysis also revealed that IQ1 could predict IQ2 significantly in all the tests (all $P < 0.001$).

Discussion After careful choice of appropriate initial test, stability of IQ in children with developmental delay was noted from pre-school through early childhood. In addition, the translated version of cognitive assessment was valid for the required context of an Asian developing country. With the current emphasis on early identification and intervention for pre-school children with developmental delay, this information bears merit in clinical practice.

Keywords autism, cognitive profile, intellectual disability, pre-school children

Correspondence: Dr For-Wey Lung, Department of Psychiatry, Kaohsiung Armed Forces General Hospital, No. 2, Zhong-zheng 1st Road, Ling-Ya District, Kaohsiung 802, Taiwan (e-mail: pincheny@gmail.com).

Introduction

Cognitive ability, as measured with standardised tests and reported as developmental quotients(DQ) or intelligence quotient (IQ), is among one of the many indexes, including academic achievement, school readiness test scores, grade retention and placement in special education, that are used to characterise child cognitive development (Anderson *et al.* 2003). IQ is considered to be stable over time in normal school-age children (Honzik *et al.* 1948; Tuma & Appelbaum 1980; Schuerger & Witt 1989; Moffitt *et al.* 1993; Neyens-Lidwien & Aldenkamp 1997) and school-age children with clinical conditions (Freeman *et al.* 1985; Lord & Schopler 1989a) as they are followed through childhood.

Regarding pre-school children with developmental disorder, which shows changes with respect to prevailing symptomatology with age, the level of cognitive impairment as revealed by IQ score has previously been identified as an early predictor of outcome (Nordin & Gillberg 1998; Szatmari 2003; Tager-Flusberg & Joseph 2003). Stability of pre-school IQs in children with developmental disorders such as intellectual disability (ID) and autism has been reported in Western developed countries (Lord & Schopler 1989b; Field *et al.* 1990; Keogh *et al.* 1997; Dietz *et al.* 2007) and Asian developing countries (Yang *et al.* 2003, 2010). One study even found that there was more stability between infant test scores and childhood IQs for children with developmental disabilities than for children without such disabilities (Sattler 2001). However, it is acknowledged that administration of cognitive tests to pre-school children with developmental delays is not easy. Clinical psychologists usually have to choose among different assessment tools so that children can be given the most chronologically age-appropriate and ability-matching tests on which they can receive a basal score. Moreover, it may be difficult to test some pre-school children with developmental delay because they have limited attention skills and/or inadequate communicative and social abilities. For children who have severe communication problems affecting both verbal and non-verbal modes, it can be hard to convey to them the requirements of a test. The impairments in social interaction may make it difficult for the examiner to establish a good rapport with the child, and the child is usually not responsive to the examiner's demands in testing. Hence, the test employed was proposed to be one of the factors related to change and stability in pre-school children with developmental disorders, and was shown to be so by several previous autism research conducted in developed countries (Harris & Handleman 2000; Magiati & Howlin 2001).

Early intervention for children with developmental delay has been shown to improve overall family functioning, child behaviour and long-term adult outcome (American Academy of Paediatrics 1994). As the current emphasis of clinical care is on early detection and intervention of pre-school children with developmental delay, more information is needed on the validity of early assessments and possible outcome, especially when some of the cognitive assessment tools used in non-Western cultures were adapted and translated from the English editions. As part of an ongoing clinical service programme in an Asian developing country, this study attempted to determine the stability of different cognitive tests applied in pre-school periods for developmentally delayed children with various diagnoses as they were followed through early childhood.

Methods

Participants

For participant enrolment, we retrospectively reviewed the medical records of pre-school children visiting the developmental clinic of an Asian developing country from the centre's inauguration date (April 1997) to December 2005. This developmental clinic is a government-designated regional referral centre for pre-school children suspected to have developmental problems. The range of this cohort was chosen so that children entering elementary school as of 1 September 2009 were selected. Among the pre-school-age children referred for evaluation of possibilities of developmental delay during this time frame, 631 pre-schoolers were ascertained to have uneven/delayed cognitive profile by formal cognitive tests in their initial pre-school visit. The child was defined as 'delayed' in cognitive development when the child's mental developmental index (MDI) or full-scale intelligence quotients (FSIQ) were less than 85. The definition of 'uneven

cognitive profile' was any of the following conditions: (1) the absolute difference between verbal intelligence quotients (VIQ) and performance intelligence quotients (PIQ) obtained from a norm-referenced standardised test was more than 15 (i.e. one standard deviation of the standardised scores) and (2) VIQ or PIQ could not be determined because the child's raw scores were zero in more than two sub-scale items. Of the 631 pre-schoolers, 398 of them came back for follow-up when they were older than 6 years of age. Of these 398 children at the follow-up, 313 participants received the Wechsler Series of Intelligence Tests – Chinese [i.e. either by Wechsler Preschool and Primary Scale of Intelligence – Revised – Chinese version (WPPSI-R; Wechsler 1989, 2000) or by Wechsler Intelligence Scale for Children-III-Chinese (WISC-III; Wechsler 1991, 1997)] for cognitive re-evaluation. The final participants for the current analysis were these 313 children defined at pre-school period as cognitively delayed or uneven. The medical chart for each eligible child was reviewed and we extracted the following data in the medical records for analysis: age, gender, medical condition, information of developmental progression, results of standardised psychometric testing and family condition. This study was approved by the Institute Review Board of the hospital.

Cognitive measures

Because this current investigation used cognitive measures from assessments performed as part of an ongoing clinical service, tests were not randomly assigned to children as fixed experimental protocol. Rather, the children were given the most chronologically age-appropriate tests on which they could receive a basal score. Cognitive assessment was carried out by one or more of the following tests in the initial pre-school assessment time (T1): (1) Bayley Scale of Infant Development – Second Edition (Bayley-II; Bayley 1993). This is a widely used measure of development for preverbal children aged 1–42 months that has well-developed norms. From Bayley-II, an MDI was derived. MDI scores are norm-referenced standard scores (mean = 100, standard deviation = 15). The MDI reflect skills in cognitive, problem-solving, language and perceptual domains, but individual scores for language and visual-motor problem-solving abilities were not reported separately. An MDI of 70 to 84 is classified as 'mildly delayed performance'. A score of not more than 69 is defined as 'significantly delayed performance'. When the child's chronological age exceeded the 42-month upper-age limit of this test, ratio DQs (age equivalent/chronological age × 100) were derived. (2) Leiter International Performance Scale – Revised (Leiter-R; Leiter 1997). The Leiter-R is designed to be a non-verbal measure of intelligence and it consists of 20 sub-tests divided between two batteries: Visualization and Reasoning Battery and an Attention and Memory Battery. The Leiter-R is for individuals between 2 years and 20 years 11 months of age. The Visualization and Reasoning Battery consists of 10 sub-tests in all; 4 sub-tests comprise a Brief IQ Screener for all ages, and two sets of 6 sub-tests are used to obtain an FSIQ. IQ and composite scores have a mean of 100 and standard deviation of 15. (3) WPPSI-R (Wechsler 1989, 2000). WPPSI-R assesses the intelligence of children between 3 and 7 years of age. Three IQs could be derived from the Wechsler Series of Intelligence Tests: FSIQ, VIQ and PIQ. Each IQ score has a mean of 100 and standard deviation of 15.

For the purpose of statistical analysis, representative cognitive measures performed at T1 and follow-up childhood period (T2) were chosen and coded as IQ1 and IQ2, respectively, by the following principles: (1) when the child could be tested by WPPSI-R at T1, FSIQ at T1 would be used as IQ1. However, when the absolute difference of scores between VIQ and PIQ at T1 was more than 15, the FSIQs were usually not reported by the assessing psychologist, and hence VIQs obtained at T1 would be coded as IQ1. When VIQ could not be determined because the child's raw scores were zero in more than two sub-tests items, the PIQ obtained at T1 would be coded as IQ1. (2) When Leiter-R was chosen as the test at T1, full-scale scores from Leiter-R would be coded as IQ1. (3) When only Bayley-II could be administered for the child at T1, MDI at T1 would be used as IQ1. When the child's chronological age exceeded the 42-month upper-age limit of the Bayley-II, DQ would be used as IQ1 instead. As for the coding of IQ2, as all the participants eligible for current investigational analysis received Wechsler Series of Intelligence Tests at T2,

FSIQ from the WISC-III or WPPSI-R administered at T2 was coded as the IQ2.

The final participant pool consisted of 313 children. Two hundred and seventy-two of them were, by definition, cognitively delayed (i.e. FSIQ or MDI less than 85) at T1. Forty-one of them were of uneven cognitive profile by the following conditions: (1) 25 children had absolute discrepancies in scores between VIQs and PIQs more than 15 (average absolute discrepancy: 30.1); (2) nine children unable to be tested by WPPSI-R were scored to be more than 85 in FSIQ by Leiter-R; and (3) seven children received zero by raw score in more than two items of the VIQ sub-tests. These final 313 children were on average 50.0 ± 11.7 months old at T1 and 88.6 ± 21.1 month old at T2. There were more boys than girls (227 vs. 86). The average interval between T1 and T2 was 38.6 ± 22.1 months. Mean IQ1 was 65.8 ± 15.4 while mean IQ2 at follow-up was 73.2 ± 17.9. For the 313 children at T1, 86 of them received Bayley-II, 28 of them received Leiter-R and 199 of them received WPPSI-R as the initial cognitive test.

Diagnostic groups

For our analysis, disease classification of the 313 children was determined through a review of data abstracted from medical records by the first author. Children met the diagnosis of autism spectrum disorder (ASD) if their records documented behaviours consistent with the Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition – Text Revision (DSM-IV-TR) criteria (American Psychiatric Association 2000) for autistic disorder, pervasive developmental disorder – not otherwise specified or Asperger disorder. Diagnostic tools of Autism Diagnostic Interview – Revised and Autism Diagnostic Observation Schedule were not used because these instruments were not available in Chinese versions before 2009. Children met the case definition of non-autistic ID if their FSIQ or DQ or full-scale Leiter-R scores were below 70 when they were aged more than 6 years; also, the review of their developmental records must exhibit lack of adequate data to support an ASD diagnosis. Children met the case definition of DSM-IV-TR mixed receptive-expressive language disorder (LD) when obvious and impairing problems in comprehension and expression of verbal communication were reported in the developmental record with special note of language difficulties in excess of those noted in cognitive problems. In addition, there had to be lack of adequate evidence in the record to support diagnosing the child as ASD. The rationale for this diagnosis by retrospective chart review is because we believe there may be an overlap in symptomatology between autism and other communication disorders and global ID in the young pre-school children. In the total 313 children investigated retrospectively by medical records, 73 (23.3%) children were diagnosed as ASD, 168 (53.7%) children were diagnosed as non-autistic ID, 58 (18.5%) children were diagnosed as LD and the remaining 14 children were of other heterogeneous diagnoses (e.g. cerebral palsy, hearing impairment, rare muscular disorder, etc.). Data from the various subgroups were subsequently analysed.

Data analysis

Stability of scores from T1 to T2 was evaluated by using the *r*-squared for Pearson's correlation coefficient for the total 313 children whose IQ assessments were initially performed by three different tests (i.e. either Bayley-II, Leiter-R or WPPSI-R) and in each of the diagnostic subgroups. For the WPPSI-Wechsler subgroup, the VIQ and PIQ stability were also explored by the *r*-squared for Pearson's correlation coefficient. Correlation coefficients (*r*) ranging from 0 to 0.25 suggested little or no relationship, 0.25 to 0.50 indicated a fair degree of relationship, 0.50 to 0.75 was moderate to good and those greater than 0.75 were good to excellent (Portney & Watkins 2000). *R*-squared for Pearson coefficients was used for data presentation because we would like to take proportion of variance overlapping between T1 IQ and T2 IQ into consideration, and the metric *r*-squared would better convey the effect size of the overlap than simply using the correlation coefficient *r*. Then, multiple regressions were used to see the predicting effect of IQ1 scores on the outcome of IQ2. To further evaluate the test–test difference, multiple regressions were analysed again with outcome of 'IQ2 minus IQ1' and predictor as IQ1. Age at initial assessment was also analysed by linear regression to see whether it has predicting effect for test–test difference. Gender and

duration of follow-up (centred at average follow-up 39 months) were adjusted in an afterwards calculation in adjusting for potential confounding variables. All data analysis was conducted using SPSS for Windows Version 15.0, and all statistical tests were performed at the two-tailed significance level of 0.05.

Results

Effect of different tests on intelligence quotient correlation and test–test difference

When the 313 children were analysed to see the correlation of IQ1 administered by different tests and IQ2, the results of r-squared for Pearson correlation coefficients indicated a pattern of moderate to good correlations in all three tests applied (all $P < 0.001$, r-squared = 0.43, 0.47 and 0.50, respectively, for IQ1 by Bayley-II, Leiter-R and WPPSI-R-Chinese). For the WPPSI-Wechsler subgroup, the r-squared correlations of VIQs at T1 and T2 and PIQs at T1 and T2 were also good (both $P < 0.001$; r-squared = 0.41 and 0.61, respectively, for VIQ and PIQ). Multiple regressions analysis revealed that IQ1 could predict IQ2 significantly in all the tests (all $P < 0.001$). Age at initial assessment also showed no statistically meaningful significance in predicting the quantity of test–test IQ difference.

When we adjusted potential confounding variables in these multiple regressions by gender and follow-up interval (centred at average follow-up of 39 months), the analysis revealed that there was meaningful statistical significance ($P < 0.05$) for the WPPSI-R group in using IQ1 to predict the test–test IQ difference. However, the parameter estimate was quite small ($\beta = -0.12$). In summary, the IQs were stable across time by three different initial tests. Results of relationships between initial and follow-up IQs in the total 313 children by different tests are reported in Table 1.

Effect of clinical diagnosis on intelligence quotient correlation

When the various diagnostic groups were analysed, the results of correlation coefficients indicated a pattern of moderate to good correlations between IQ1 and IQ2 in the ASD subgroup (r-squared = 0.26–0.72, all $P < 0.05$). In the ID subgroup, the correlation of IQ1 and IQ2 was fair for the Bayley-II tested (r-squared = 0.22, $P < 0.001$) and good for the WPPSI-R tested (r-squared = 0.42, $P < 0.001$). In the LD subgroup, the WPPSI-R tested pre-schooler had fair correlations between IQ1 and IQ2 (r-squared = 0.16, $P = 0.01$); while the Bayley-II tested and Leiter-R tested were found to have no correlation between initial IQ and

Table 1 Relationships between intelligence quotient (IQ) scores for assessments followed through early childhood in 313 pre-school children with uneven/delayed cognition

		Mean age (months)		Mean IQ (mean ± SD)		I	II	III
Initial test	*n*	T1	T2	IQ1	IQ2	r-squared for Pearson coefficient	β	β
Bayley-II	86	38	86	57.24 ± 11.76	62.67 ± 16.94	0.43***	0.95***	−0.05
Leiter-R	28	52	81	87.86 ± 18.41	78.46 ± 19.37	0.47***	0.73***	−0.28
WPPSI-R	199	54	89	66.38 ± 12.94	76.95 ± 16.18	0.50***	0.88***	−0.12

T1: age in months at initial assessment; T2: age in months at follow-up.

IQ1: initial pre-school IQ; IQ2: follow-up childhood IQ.

I: r-squared for Pearson correlation between IQ1 and IQ2.

II: Multiple linear regression using IQ2 as outcome variable, IQ1 as predictor.

III: Multiple linear regression using test–test difference as outcome variable, IQ1 as predictor.

* $P < 0.05$; *** $P < 0.001$; β: parameter estimate.

Bayley-II: Bayley Scale of Infant Development – Second Edition.

Leiter-R: Leiter International Performance Scale – Revised.

WPPSI-R: Wechsler Preschool and Primary Scale of Intelligence – Revised.

Table 2 Relationships between intelligence quotient (IQ) scores for assessments followed through early childhood in diagnostic subgroups of 313 pre-school children with uneven/delayed cognitive development

		Mean age (years-months)		Mean IQ (mean ± SD)		*r*-squared for Pearson coefficient	*P*-value
	n	T1	T2	T1	T2		
ASD ($n = 73$)							
Bayley-II/Wechsler	20	3-4	7-1	58.70 ± 9.81	67.95 ± 19.06	0.26	0.023
Leiter-R/Wechsler	16	4-5	6-8	89.88 ± 19.19	79.44 ± 21.71	0.72	<0.001
WPPSI-R/Wechsler	37	4-7	7-2	72.76 ± 13.17	85.11 ± 15.13	0.37	<0.001
ID ($n = 168$)							
Bayley-II/Wechsler	56	3-2	7-4	53.12 ± 9.22	56.27 ± 11.06	0.22	<0.001
Leiter-R/Wechsler	7	4-3	6-7	75.14 ± 12.86	68.00 ± 10.46	0.16	0.38
WPPSI-R/Wechsler	105	4-6	7-8	59.96 ± 10.88	68.43 ± 13.94	0.42	<0.001
LD ($n = 58$)							
Bayley-II/Wechsler	5	2-8	6-8	75.20 ± 4.32	89.40 ± 17.53	0.03	0.78
Leiter-R/Wechsler	5	4-1	7-8	99.20 ± 14.04	90.00 ± 15.44	0.00	0.96
WPPSI-R/Wechsler	48	4-5	7-2	72.67 ± 9.73	87.29 ± 11.96	0.16	0.01
Other ($n = 14$)							
Bayley-II/Wechsler	5	2-3	6-0	79.60 ± 4.04	86.60 ± 8.02	0.20	0.46
WPPSI-R/Wechsler	9	4-4	7-1	81.44 ± 6.93	87.78 ± 6.74	0.29	0.14

ASD, Autism spectrum disorder; ID, non-autistic intellectual disability; LD, mixed receptive-expressive language disorder.
Bayley-II: Bayley Scale of Infant Development – Second Edition.
Leiter-R: Leiter International Performance Scale – Revised.
WPPSI-R: Wechsler Preschool and Primary Scale of Intelligence – Revised.
Wechsler: WPPSI-R or Wechsler Intelligence Scale for Children-III.

follow-up IQ, although the calculation had no statistical significance. Results of correlations between initial and follow-up IQs in the diagnostic subgroups are reported in Table 2. It was noted that the numbers of participants for the Leiter-R tested ID pre-schoolers ($n = 7$), Leiter-R tested LD pre-schoolers ($n = 5$) and Bayley-II tested LD pre-schoolers ($n = 5$) were small. The non-significant results obtained in statistical analysis were probably best explained by the limited sample size.

Discussion

Our main results showed the IQ scores obtained from three frequently applied standardised cognitive tests in the context of an Asian developing country were stable for pre-schoolers of developmental disorders as they were followed through early childhood. These findings of stability of IQ in children with developmental delays were in agreement with those previously reported in children with intellectual deficit, autism and developmental language disorders from Western developed countries (Lord & Schopler 1989b; Keogh *et al.* 1997; Clegg *et al.* 2005; Dietz *et al.* 2007; Begovac *et al.* 2009). Previous Western studies have also reported age of initial assessment to be an important factor in the exploration of stability of pre-school testing for certain subgroups of children such as autism. Cognitive testing of autistic pre-schoolers older than 4 years of age was reported to give more stable measures compared with the testing of younger children of autism (Lord & Schopler 1989a,b). We did not group our participants (range from 13-month-olds to 64-month-olds at T1) by different age of initial test for analysis because we considered the groups thus constituted would be of heterogeneous diagnoses and render the interpretation difficult. Nevertheless, linear regression showed that age at initial assessment did not have predictive effect for test–test IQ difference for the whole group of developmentally delayed children with various clinical diagnoses.

Another important message from the current survey was the validity of the translated Chinese

version of the Wechsler Intelligence Test series for the clinical population. These tests have already been applied extensively in educational settings since their publication. A number of validity studies have already appeared in the literature, examining issues such as factorial validity and short-form validity (Chen *et al.* 2000, 2002; Chen & Zhu 2009). We believe our current investigation may provide additional evidence of the validity of the Chinese version of the Wechsler series of intelligence assessment. For future studies, it would be of interest to examine the predictive validity of these developmental tests through such as examining the prediction of measured achievement at the end of the first-grade school from pre-school cognitive tests as prediction of academic success is arguably the most important criterion of validity for any IQ assessment.

The IQ stability in the majority of diagnostic subgroups by different tests was fair to good. It was only in the subgroup of mixed receptive-expressive LD that the Bayley-II tested and Leiter-R tested scores were noted to have lack of stability between initial IQ and follow-up IQ. Although being of no statistical significance, these results were probably best explained by the limited sample size. In terms of clinical relevance, the importance of the statistical analysis may be less meaningful, although the finding still deserves attention for this subgroup of children with mixed receptive-expressive LD. Language disorder is a highly heterogeneous clinical condition in which the defining symptoms vary a lot depending upon age and classification system. Language problems can involve difficulty with grammar (syntax), words or vocabulary (semantics), the rules and system for speech sound production (phonology), units of word meaning (morphology) and the use of language particularly in social contexts (pragmatics). The diagnostic system we used, that is, DSM-IV-TR, coded these clinical conditions as 'communication disorders'; however, there were also several different kinds of classification systems with various terminology (American Speech-Language-Hearing Association 1993). Pre-school language difficulties are frequently the precursors of language and academic difficulties that persist throughout childhood and adolescence; however, there is substantial variation in outcome that is not fully understood (Stothard *et al.* 1998). In the current study, the data we analysed were from children with mixed receptive-expressive LDs, which are supposedly of the most severe sub-type of language disorders, and our result brought more questions than it might answer. Future studies with increased sample size will be needed in delineating the developmental course and prognosis of mixed receptive-expressive LD.

No test is able to measure all abilities that lie within the complex domain of intelligence. Our current tests measure only a segment of the diverse abilities that define Howard Gardner's theory of intelligence (Gardner 1993) and assess only one dimension of Robert Sternberg's triarchic theory, namely, analytical abilities, but not creative or practical abilities (Sternberg 1984), yet they do measure abilities that are able to be assessed objectively and are shown to be stable over time by our results. However, overall IQ scores may mask the complex nature of cognitive abilities in special subgroup of children, who may show similar overall performance to a comparison group on a specific task, but may complete the task in an entirely different manner. It should be of note that recent research has suggested that it might be useful to evaluate cognition from a process versus outcome approach (Volkmar *et al.* 2004).

Several limitations of our current investigation deserve further attention. First, this study was clinically based and as such, it has a limited generalisability than if the sample had been drawn from the community. Second, we only recruited children who could be tested by the Wechsler series of intelligence tests at the follow-up. As children were given the most chronologically age-appropriate tests, children who continued to be given the same pre-school test when they were no longer pre-school children were those who were more delayed cognitively. Hence, the stability reported here was likely because of the selective exclusion of children who would not have scored at all on the WISC-III or WPPSI-R. The third major limitation was our inability to adjust for treatment effects so that test–treatment interactions could not be rejected. A strong body of evidence shows that early childhood development programmes have a positive effect on preventing delay of cognitive development and increasing readiness to learn (Anderson *et al.* 2003). Many behaviours associated with better test perfor-

mance, such as attention, cooperation and motivation, are targeted in early intervention programmes; however, the services that our participants received were not based on theories of education or treatment developed for this longitudinal investigation, but rather on several different approaches and personal experiences of the pre-school staff from various educational settings. As a result, it is not possible for us to determine the complete picture of intervention each individual child had received from retrospective medical chart review.

The conclusion drawn from our analysis is that after careful choice of appropriate initial testing, stability of IQ in children with developmental delay was noted from pre-school through early childhood. With the emphasis on early identification and intervention for pre-school children with developmental delay, this information bears merit in clinical practice.

Acknowledgements

The authors thank the Statistical Analysis Laboratory, Department of Medical Research, Kaohsiung Medical University Hospital, Kaohsiung Medical University for their help.

References

American Academy of Paediatrics (1994) Screening infants and young children for developmental disabilities. American academy of paediatrics committee on children with disabilities. *Pediatrics* **93,** 863–5.

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision.* American Psychiatric Association, Washington, DC.

American Speech-Language-Hearing Association (ASHA) (1993) Definitions of communication disorders and variations. Ad hoc committee on service delivery in the schools. *ASHA. Supplement* **35,** 40–1.

Anderson L. M., Shinn C., Fullilove M. T., Scrimshaw S. C., Fielding J. E., Normand J. *et al.* (2003) The effectiveness of early childhood development programs. A systematic review. *American Journal of Preventive Medicine* **24** (3 Suppl.), 32–46.

Bayley N. (1993) *Manual for The Bayley Scales of Infant Development, Second Edition.* Psychological Corporation, New York.

Begovac I., Begovac B., Majic G. & Vidovic V. (2009) Longitudinal studies of IQ stability in children with childhood autism – literature survey. *Psychiatria Danubina* **21,** 310–19.

Chen H. & Zhu J. (2009) Testing for WISC-III factorial invariance across gender. *[Psychological Testing] Ce Yan Nian Kan* **56,** 1–18.

Chen H. Y., Chen J. H. & Zhu J. J. (2000) Factor structure and variance partitionment of the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) in Taiwan. *[Psychological Testing] Ce Yan Nian Kan* **47,** 17–33.

Chen H. Y., Yang T. R., Zhu J. & Chang B. S. (2002) Validities of the WISC-III short forms based on a multiform/multi-sample/multi-method design. *[Psychological Testing] Ce Yan Nian Kan* **49,** 155–82.

Clegg J., Hollis C., Mawhood L. & Rutter M. (2005) Developmental language disorders – a follow-up in later adult life. Cognitive, language and psychosocial outcomes. *Journal of Child Psychology and Psychiatry* **46,** 128–49.

Dietz C., Swinkels S. H., Buitelaar J. K., Van Daalen E. & Van Engeland H. (2007) Stability and change of IQ scores in preschool children diagnosed with autistic spectrum disorder. *European Child & Adolescent Psychiatry* **16,** 405–10.

Field M., Fox N. & Radcliffe J. (1990) Predicting IQ change in preschoolers with developmental delays. *Journal of Developmental and Behavioral Pediatrics* **11,** 184–9.

Freeman B. J., Ritvo E. R., Needleman R. & Yokota A. (1985) The stability of cognitive and linguistic parameters in autism: a five-year prospective study. *Journal of the American Academy of Child and Adolescent Psychiatry* **24,** 459–64.

Gardner H. (1993) *Frames of Mind: The Theory of Multiple Intelligence.* Basic Books, New York.

Harris S. L. & Handleman J. S. (2000) Age and IQ at intake as predictors of placement for young children with autism: a four- to six-year follow-up. *Journal of Autism and Developmental Disorders* **30,** 137–42.

Honzik M. P., MacFarlane J. W. & Allen L. (1948) The stability of mental test performance between two and eighteen years. *The Journal of Experimental Education* **17,** 309–24.

Keogh B. K., Bernheimer L. P. & Guthrie D. (1997) Stability and change over time in cognitive level of children with delays. *American Journal on Mental Retardation* **101,** 365–73.

Leiter R. G. (1997) *Leiter International Performance Scale – Revised.* Stoelting, Chicago, IL.

Lord C. & Schopler E. (1989a) The role of age at assessment, developmental level, and test in the stability of intelligence scores in young autistic children. *Journal of Autism and Developmental Disorders* **19,** 483–99.

Lord C. & Schopler E. (1989b) Stability of assessment results of autistic and non-autistic language-impaired children from preschool years to early school age. *Journal of Child Psychology and Psychiatry* **30,** 575–90.

Magiati I. & Howlin P. (2001) Monitoring the progress of preschool children with autism enrolled in early intervention programmes: problems in cognitive assessment. *Autism* **5,** 399–406.

Moffitt T. E., Caspi A., Harkness A. R. & Silva P. A. (1993) The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry* **34,** 455–506.

Neyens-Lidwien G. J. & Aldenkamp A. P. (1997) Stability of cognitive measures in children of average ability. *Child Neuropsychology* **3,** 161–70.

Nordin V. & Gillberg C. (1998) The long-term course of autistic disorders: update on follow-up studies. *Acta Psychiatrica Scandinavica* **97,** 99–108.

Portney L. G. & Watkins M. P. (2000) *Foundations of Clinical Research: Applications to Practice*, 2nd edn. Prentice Hall Health, Upper Saddle River, NJ.

Sattler J. M. (2001) *Assessment of Children: Cognitive Applications*, 4th edn. Jerome M. Sattler Publisher Inc., San Diego, CA.

Schuerger J. M. & Witt A. C. (1989) The temporal stability of individually tested intelligence. *Journal of Clinical Psychology* **45,** 294–302.

Sternberg R. J. (1984) *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge University Press, Cambridge.

Stothard S. E., Snowling M. J., Bishop D. V., Chipchase B. B. & Kaplan C. A. (1998) Language-impaired preschoolers: a follow-up into adolescence. *Journal of Speech, Language, and Hearing Research* **41,** 407–18.

Szatmari P. (2003) *Outcome in autism spectrum disorders*. Paper presented at the Social Brain Conference, Gothenburg, Sweden.

Tager-Flusberg H. & Joseph R. M. (2003) Identifying neurocognitive phenotypes in autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **358,** 303–14.

Tuma J. & Appelbaum A. S. (1980) Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement* **40,** 671–8.

Volkmar F. R., Lord C., Bailey A., Schultz R. T. & Klin A. (2004) Autism and pervasive developmental disorders. *Journal of Child Psychology and Psychiatry* **45,** 135–70.

Wechsler D. (1989) *Manual for the Wechsler Preschool and Primary Scale of Intelligence – Revised*. Psychological Corp., New York.

Wechsler D. (1991) *Manual for the Wechsler Intelligence Scale for Children, Third Edition*. Psychological Corp., New York.

Wechsler D. (1997) *Manual for the Wechsler Intelligence Scale for Children, Third Edition*, Chinese version. Chinese Behavioural Science Corporation, Taipei.

Wechsler D. (2000) *Manual for the Wechsler Preschool and Primary Scale of Intelligence, Revised*, Chinese version. Chinese Behavioural Science Corporation, Taipei.

Yang P., Jong Y. J., Hsu H. Y. & Chen C. S. (2003) Preschool children with autism spectrum disorders in Taiwan: follow-up of cognitive assessment to early school age. *Brain and Development* **25,** 549–54.

Yang P., Lung F. W., Jong Y. J., Hsu H. Y. & Chen C. C. (2010) Stability and change of cognitive attributes in children with uneven/delayed cognitive development from preschool through childhood. *Research in Developmental Disabilities* **31,** 895–902.

Accepted 27 January 2011

Journal of Clinical Child & Adolescent Psychology, *42*(1), 139–159, 2013

ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374416.2012.736358

Routledge
Taylor & Francis Group

FUTURE DIRECTIONS

Future Directions in Psychological Assessment: Combining Evidence-Based Medicine Innovations with Psychology's Historical Strengths to Enhance Utility

Eric A. Youngstrom
Departments of Psychology and Psychiatry, University of North Carolina at Chapel Hill

Assessment has been a historical strength of psychology, with sophisticated traditions of measurement, psychometrics, and theoretical underpinnings. However, training, reimbursement, and utilization of psychological assessment have been eroded in many settings. Evidence-based medicine (EBM) offers a different perspective on evaluation that complements traditional strengths of psychological assessment. EBM ties assessment directly to clinical decision making about the individual, uses simplified Bayesian methods explicitly to integrate assessment data, and solicits patient preferences as part of the decision-making process. Combining the EBM perspective with psychological assessment creates a hybrid approach that is more client centered, and it defines a set of applied research topics that are highly clinically relevant. This article offers a sequence of a dozen facets of the revised assessment process, along with examples of corollary research studies. An eclectic integration of EBM and evidence-based assessment generates a powerful hybrid that is likely to have broad applicability within clinical psychology and enhance the utility of psychological assessments.

What if we no longer performed psychological assessment? Although assessment has been a core skill and a way of conceptualizing individual differences central to psychology, training and reimbursement have eroded over a period of decades (Merenda, 2007b). Insurance companies question whether they need to reimburse for psychological assessment (Cashel, 2002; Piotrowski, 1999). Educational systems have moved away from using ability-achievement discrepancies as a way of identifying learning disability and decreased the emphasis on individual standardized tests for individual placement (Fletcher, Francis, Morris, & Lyon, 2005). Several traditional approaches to personality assessment, such as the various interpretive systems for the Rorschach, have had their validity challenged repeatedly (cf. Meyer & Handler, 1997; Wood, Nezworski, & Stejskal, 1996). Many graduate-level training programs are reducing their emphasis on aspects of assessment (Belter & Piotrowski, 2001; Childs & Eyde, 2002; Stedman, Hatch, & Schoenfeld, 2001) and psychometrics (Borsboom, 2006; Merenda, 2007a) in their curricula, and few undergraduate programs offer courses focused on assessment or measurement. Efforts to defend assessment have been sometimes disorganized and tepid, or hampered by a lack of data even when committed and scholarly (Meyer et al., 1998).

Is this intrinsically a bad thing? Training programs, systems of care, and providers all have limited resources. Assessment might be a luxury in which some could afford to indulge, paying for extensive evaluations as a way to gain insight into themselves. However, arguments defending assessment as a major clinical activity need to appeal to utility to be persuasive (Hayes, Nelson, & Jarrett, 1987). Here, "utility" refers to adding value to individual care, where the benefits deriving from the assessment procedure clearly outweigh the costs, even when the costs combine fiscal expense with other

Thanks to Guillermo Perez Algorta for comments and suggestions
Correspondence should be addressed to Eric A. Youngstrom, Department of Psychology, University of North Carolina at Chapel Hill, CB #3270, Davie Hall, Chapel Hill, NC 27599-3270. E-mail: eay@unc.edu

factors such as time and the potential for harm (Garb, 1998; Kraemer, 1992; Straus, Glasziou, Richardson, & Haynes, 2011). Although utility has often been described in contexts of dichotomous decision making, such as initiating a treatment or not, or making a diagnosis or not, it also applies to situations with ordered categories or continuous variables. Conventional psychometric concepts such as reliability and validity are prerequisites for utility, but they do not guarantee it. Traditional evaluations of psychological testing have not formally incorporated the concept of costs in either sense—fiscal or risk of harm.

Using utility as an organizing principle has radical implications for the teaching and practice of assessment. Assessment methods can justify their place training and practice if they clearly address at least one aspect of prediction, prescription, or process—the "Three Ps" of assessment utility (Youngstrom, 2008). *Prediction* refers to association with a criterion of importance, which could be a diagnosis, but also could be another category of interest, such as adolescent pregnancy, psychiatric hospitalization, forensic recidivism, graduation from high school, or suicide attempt. For our purposes, the criterion could be continuous or categorical, and the temporal relationship could be contemporaneous or prospective. The goal is to demonstrate predictive validity for the assessment procedure by any of these methods and to make a compelling case that the effect size and cost/benefit ratio suggest utility. *Prescription* refers more narrowly to the assessment providing information that changes the choice of treatment, either via matching treatment to a particular diagnosis or by identifying a moderator of treatment. Similarly, *process* refers to variables that inform about progress over the course of treatment and quantify meaningful outcomes. These could include mediating variables, or be measures of adherence or treatment response. Each of the Three Ps demonstrates a connection to prognosis and treatment. These are not the only purposes that could be served by psychological assessment, but they are some of the most persuasive in terms of satisfying stakeholders that the assessment method is adding value to the clinical process (Meehl, 1997). Many of the other conventional goals of psychological assessment (Sattler, 2002) can be recast in terms of the Three Ps and utility: Using assessment as a way of establishing developmental history or baseline functioning may have predictive value or help with treatment selection, as can assessment of personality (Harkness & Lilienfeld, 1997). Case formulation speaks directly to the process of working effectively with the individual. Gathering history for its own sake is much less compelling than linking the findings to treatment and prognosis (Hunsley & Mash, 2007; Nelson-Gray, 2003).

It was surprising to me as an educator and a psychologist how few of the commonly taught or used techniques can demonstrate any aspect of prediction, prescription, or process—let alone at a clinically significant level (Hunsley & Mash, 2007). Surveys canvassing the content of training programs at the doctoral and internship level (Childs & Eyde, 2002; Stedman et al., 2001; Stedman, Hatch, Schoenfeld, & Keilin, 2005), as well as evaluating what methods are typically used by practicing clinicians (Camara, Nathan, & Puente, 1998; Cashel, 2002), show that people tend to practice similar to how they were trained. There is also a striking amount of inertia in the lists, which have remained mostly stable for three decades (Childs & Eyde, 2002). Content has been set by habits of training, and these in turn have dictated habits of practice that change slowly if at all.

When I first taught assessment, I used the courses I had taken as a graduate student as a template and made some modifications after asking to see syllabi from a few colleagues. The result was a good, conventional course; but the skills that I taught had little connection to the things that I did in my clinical practice as I pursued licensure. Much of my research has focused on assessment, but that created a sense of cognitive dissonance compared to my teaching and practice. One line of research challenged the clinical practice of interpreting factor and subtest scores on cognitive ability tests. These studies repeatedly found little or no incremental validity in more complicated interpretive models (e.g., Glutting, Youngstrom, Ward, Ward, & Hale, 1997), yet they remained entrenched in practice and training (Watkins, 2000). The more disquieting realization, though, was that my own research into assessment methods was disconnected from my clinical work. If conventional group-based statistics were not changing my own practice, why would I put forth my research to students or to other practitioners? Why was I not using the assessments I taught in class? When I reflected on the curriculum, I realized that I was teaching the "same old" tests out of convention, or out of concern that the students needed to demonstrate a certain degree of proficiency with a variety of methods in order to match at a good internship (Stedman et al., 2001).

What was missing was a clear indication of utility for the client. Reviewing my syllabi, or perusing any of the tables ranking the most popular assessment methods, emphasized the disconnect: Does scoring in a certain range on the Wechsler tests make one a better or worse candidate for cognitive behavioral therapy? Does verbal ability moderate response to therapies teaching communication skills? How does the Bender Gestalt test do at predicting important criteria? Do poor scores on it prescribe a change in psychological intervention? . . . or or tell about the process of working with a client? . . . - What about Draw a Person? Our most widely used tools do not have a literature establishing their validity in terms of individual prognosis or treatment, and viewed

through the lens of utility they look superfluous. Yet these are all in the top 10 most widely used for assessing psychopathology in youths, according to practitioner surveys (Camara et al., 1998; Cashel, 2002), even though they do not feature prominently in evidence-based assessment recommendations (Mash & Hunsley, 2005).

Evidence-based medicine (EBM) is rooted in a different tradition, grounded in medical decision making and initially advocated by internal medicine and other specialties bearing little resemblance to the field of psychology (Guyatt & Rennie, 2002; Straus et al., 2011). EBM has grown rapidly, however, and it has a variety of strengths that could reinvigorate psychological assessment practices if there were a way to hybridize the two traditions (Bauer, 2007). The principles of emphasizing evidence, and integrating nomothetic data with clinical expertise and patient preferences, are consistent with the goals of "evidence-based practice" (EBP) in psychology (Spengler, Strohmer, Dixon, & Shivy, 1995; Spring, 2007). Indeed, the American Psychological Association (2005) issued a statement endorsing EBP along the lines articulated by Sackett and colleagues and the Institute of Medicine. However, this is more agreement about a vision; and there is a fair amount of work involved in completing the merger of the different professional traditions. In much of what follows, I refer to EBM instead of EBP when talking about assessment, because EBM has assessment-related concepts that have not yet been discussed or assimilated in EBP in psychology. Key components include a focus on making decisions about individual cases, and knowing when there is enough information to consider something "ruled out" of further consideration or "ruled in" as a focus of treatment. EBM also has a radical emphasis on staying connected to the research literature, including such advice as "burn your textbooks—they are out of date as soon as they are published" (Straus et al., 2011). The emphasis on scientific evidence as guiding clinical practice seems philosophically compatible with the Boulder Model of training, and resonates with recent calls to further emphasize the scientific components of clinical psychology (McFall, 1991).

EBM's focus on relevance to the individual puts utility at the forefront: Each piece of evidence needs to demonstrate that it is valid and that it has the potential to help the patient (Jaeschke, Guyatt, & Sackett, 1994). However, most discussions of EBP in psychology have focused on therapy, with less explication of the concepts of evidence-based assessment (see Mash & Hunsley, 2005, for comment). Despite the shared vision of EBM and the American Psychological Association's endorsement of EBP, most of the techniques and concepts involved in assessment remained in distinct silos. For example, the terms "diagnostic likelihood ratio," "predictive power," "wait-test" or "test-treat threshold," or even "sensitivity" or "specificity" are not included as index terms in the current edition of *Assessment of Children and Adolescents* (Mash & Barkley, 2007; these terms are defined in the assessment context later in this article). A hand search of the volume found five entries in 866 pages that mentioned receiver operating characteristic analysis or diagnostic sensitivity or specificity (excluding the chapter on pediatric bipolar disorder, which was heavily influenced by the EBM approach). Of those five, one was a passing mention of poor sensitivity for an autism screener, and the other four were the exceptions among a set of 77 trauma measures reviewed in a detailed appendix. Discussions of evidence-based assessment have focused on reliability and classical concepts of psychometric validity but not application to individual decision making in the ways EBM proposes (Hunsley & Mash, 2005; Mash & Hunsley, 2005).

Conversely, treatments of EBM barely mention reliability and are devoid of psychometric concepts such as latent variables, measurement models, or differential item functioning (Guyatt & Rennie, 2002; Straus et al., 2011), despite the fact that these methods are clearly relevant to situations where the "gold standard" criterion diagnosis is missing or flawed (Borsboom, 2008; Kraemer, 1992; Pepe, 2003). Similarly, differential item functioning, tests of structural invariance, and the frameworks developed for testing statistical moderation would advance EBM's stated goals of understanding the factors that change whether the research findings apply to the individual patient (i.e., what are the moderating factors?; Cohen, Cohen, West, & Aiken, 2003) and understanding the process of change (i.e., the mediating variables; MacKinnon, Fairchild, & Fritz, 2007).

The two traditions have much to offer each other (Bauer, 2007). Because the guiding visions are congruent, it is often straightforward to transfer ideas and techniques between the EBM and psychological assessment EBP silos. The ideas from EBM have reshaped how I approach research on assessment, and reorganized my research and teaching to have greater relevance to individual cases. Our group has mostly applied these principles to the assessment of bipolar disorder (e.g., Youngstrom, 2007; Youngstrom et al., 2004; Youngstrom, Freeman, & Jenkins, 2009), but the concepts are far more broad. In the next section I lay out the approach to assessment as a general model and discuss the links to both EBM and traditional psychological assessment. This is not an introduction to EBM; there are comprehensive resources available (Guyatt & Rennie, 2002; Straus et al., 2011). Instead, I briefly describe some of the central features from the EBM approach to assessment and then lay out a sequence of steps for integrating these ideas with clinical psychology research and practice. The synthesis defines a set of new research questions and methods that are highly clinically

relevant, and it reorganizes assessment practice in a way that is pragmatic and patient focused (Bauer, 2007). The combination of EBM and psychological assessment also directly addresses the "utility gap" in current assessment practice and training (Hunsley & Mash, 2007). Sections describing research are oriented toward filling existing gaps, not reinforcing any bifurcation of research from practice.

A BRIEF OVERVIEW OF ASSESSMENT IN EBM

EBM focuses on shaping clinical ambiguity into answerable questions and then conducting rapid and focused searches to identify information that addresses each question (Straus et al., 2011). Rather than asking, "What is the diagnosis?" an EBM approach would refine the question to something like, "What information would help rule in or rule out a diagnosis of attention deficit/hyperactivity disorder (ADHD) for this case?" EBM references spend little time talking about reliability and almost no space devoted to traditional psychometrics such as factor analyses or classical descriptions of validity (cf. Borsboom, 2006; Messick, 1995). Instead, they concentrate on a Bayesian approach to interpreting tests, at least with regard to activities such as screening, diagnosis, and forecasting possible harm. The core method involves estimating the probability that a patient has a particular diagnosis, or will engage in a behavior of interest (such as relapse, recidivism, or self-injury), and then using Bayesian methods to combine that prior probability with new information from risk factors, protective factors, or test results to revise the estimate until the revised probability is low enough to consider the issue functionally "ruled out," or high enough to establish the issue as a clear target for treatment (Straus et al., 2011).

Bayes' Theorem, a way of combining probabilities, is literally centuries old (Bayes & Price, 1763). There are two ways of interpreting Bayes' Theorem: A Bayesian interpretation focuses on the degree to which new evidence should rationally change one's degree of belief, whereas a frequentist interpretation connects the inverse probabilities of two events, formally expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

In this formula, P(A) is the prior probability of the condition, before knowing the assessment result; P(A|B) is the posterior probability, or the revised probability taking into account the information value of the assessment result; and P(B|A)/P(B) conveys the degree of support that the assessment result provides for the condition, by comparing the probability of observing the result within the subset of those that have the condition, P(B|A), to the overall rate of the assessment result, P(B). For example, if 20% of the cases coming to a clinical practice have depression—base rate = P(A) = 20%—and the client scores high on a test with 90% diagnostic sensitivity to depression—P(B|A) = 90%, %, or 90% of cases with depression scoring positive—then Bayes' Theorem would combine these two numbers with the rate of positive test results regardless of diagnosis to generate the probability that the client has depression conditional upon the positive test result. If 30% of cases score positive on the test regardless of diagnosis (what Kraemer, 1992, called the "level" of the test, to distinguish it from the false alarm rate), then the probability that the client has depression rises to 60%. Conversely, if the client had scored below threshold on the same test, then the probability of depression drops to less than 3%. The example shows the potential power of directly applying the test results to the individual case but also illustrates the difficulty of combining the information intuitively, as well as the effort involved in traditional implementations of the Bayesian approach.

Luminaries in clinical psychology such as Paul Meehl (Meehl & Rosen, 1955), Robyn Dawes (Dawes, Faust, & Meehl, 1989), and Dick McFall (McFall & Treat, 1999) have advocated incorporating it into everyday clinical practice. Some practical obstacles have delayed the widespread adoption of the method, including that it requires multiple steps and some algebra to combine the information, and the posterior probability is heavily dependent on the base rate of the condition. An innovation of the EBM approach is to address these challenges by offering online calculators or a "slide rule" visual approximation, a probability nomogram (see Figure 1), avoiding the need for computation, albeit at the price of some loss in precision (Straus et al., 2011). The nonlinear spacing of the markers on each line geometrically accomplishes the same effect as transforming prior probabilities (the left-hand line of the nomogram) into odds, then multiplying by the change in the diagnostic likelihood (plotted on the center line) to extrapolate to the posterior probability (the right-hand line), again avoiding the algebra to convert the posterior odds back into a probability (see the appendix, or Jenkins, Youngstrom, Washburn, & Youngstrom 2011, for a worked illustration).

A second, more conceptual innovation developed by EBM is to move past dichotomous "positive test/negative test result" thinking and to suggest a multitiered way of mapping probability estimates onto clinical decision making. In theory, the probability estimate of a target condition could range from 0% to 100% for any given case. In practice, almost no cases would have estimated probabilities of exactly 0% or 100%, and few might even get close to those extremes given the limits of currently available assessment methods. The pragmatic insight is that we do not need such

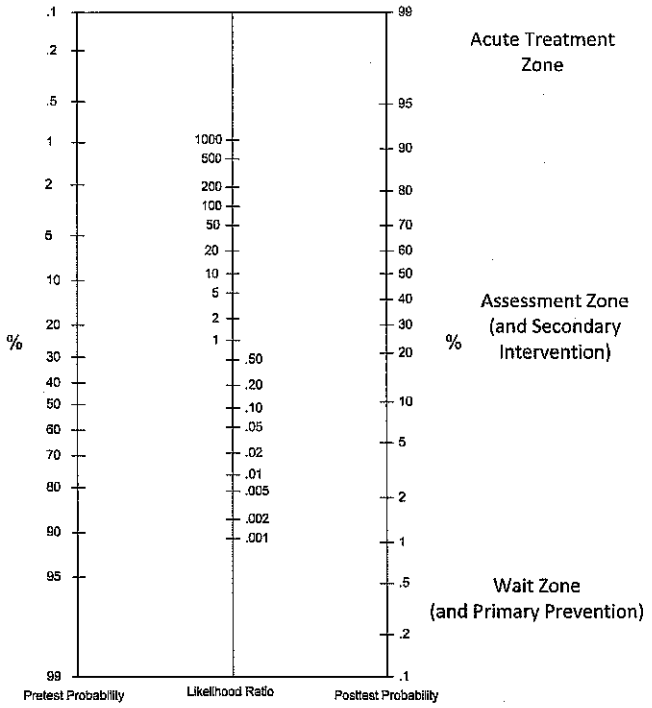

FIGURE 1 Probability nomogram for combining probability with likelihood ratios. *Note:* Straus et al. (2011) provided the rationale and examples of using the nomogram. Jenkins et al. (2011) illustrated using it with a case of possible pediatric bipolar disorder, and Frazier and Youngstrom (2006) with possible attention deficit/hyperactivity disorder.

extreme probability levels in order to make most clinical decisions (Straus et al., 2011). If the revised probability is high enough, then it makes sense to initiate treatment, in the same way that if the weather forecast calls for a 95% chance of showers, then we would do well to dress for rain. EBM calls the threshold where it makes sense to initiate treatment the "test-treat threshold"—probabilities above that level indicate intervention, whereas below that same point suggest continued assessment (Straus et al., 2011). Similarly, there is a point where the probability is sufficiently low to consider the target condition "ruled out" even though the probability is not zero. Below this "wait-test" threshold, EBM argues that there is no utility in continued assessment, nor should treatment be initiated. The two thresholds divide the range of probabilities and map them onto three clinical actions: actively treat, continue assessing, or decide that the initial hypothesis is not supported—and either assess or treat other issues (Guyatt & Rennie, 2002; Straus et al., 2011).

A third innovation in EBM is not to specify the exact locations for the wait-test and test-treat thresholds a priori. Instead, EBM provides a framework for incorporating the costs and benefits attached to the diagnosis, the test, and the treatment, and then using them to help decide where to set the bars for a particular case (Straus et al., 2011). Even better, there are ways of engaging the patient and soliciting personal preferences, including them in the decision-making process. For effective, low-risk, low-cost interventions, the treatment threshold might be so low that it makes sense to skip the assessment process entirely, as happens with routine vaccinations, or with the addition of fluoride to drinking water (Youngstrom, 2008). Conversely, for clinical issues where the treatment is freighted with risks, it makes sense to reserve the intervention until the probability of the target diagnosis is extremely high. For many families, atypical antipsychotics may fall in that category, given the serious side effects and the relative paucity of information about long-term effects on development (Correll, 2008). The EBM method creates a process for collaboratively weighing the costs, benefits, and preferences. This has the potential to empower the patient and customize treatment according to key factors, and it moves decision making from a simple, dichotomous mode to much more nuanced gradations. For the same patient, the test-treat thresholds might be more stringent for initiating medication than therapy, and so based on the same evidence it may make sense to start therapy, and wait to decide about medication until after additional assessment data are integrated.

These three innovations of (a) simplifying the estimation of posterior probabilities; (b) mapping the probability onto the next clinical action; and (c) incorporating the risks, benefits, and patient preferences in the decision-making process combine to restructure the process of assessment selection and interpretation. Assimilating these ideas has led to a multistep model for evaluating potential pediatric bipolar disorder (Youngstrom, Jenkins, Jensen-Doss, & Youngstrom, 2012). This model starts with estimates of the rate of bipolar in different settings, combines that with evidence of risk factors such as familial history of bipolar disorder, and then adds test results from either the Achenbach (Achenbach & Rescorla, 2001) or more specialized mood measures. Our group has published some of the needed components, such as the "diagnostic likelihood ratios" (DLRs; Straus et al., 2011) that simplify using a probability nomogram (Youngstrom et al., 2004), and vignettes illustrating how to combine test results and risk factors for individual cases (Youngstrom & Duax, 2005; Youngstrom & Kogos Youngstrom, 2005). We have tested whether weights developed in one sample generalize to other demographically and clinically different settings (Jenkins, Youngstrom, Youngstrom, Feeny, & Findling, 2012). These methods have large effects on how practicing clinicians interpret information, making their estimates more accurate and consistent, and eliminating a tendency to overestimate the risk of bipolar disorder (Jenkins, et al., 2011).

The methods are not specific to bipolar disorder: The core ideas were developed in internal medicine and have generalized throughout other medical practices (Gray, 2004; Guyatt & Rennie, 2002). These ideas define a set of clinically relevant research projects for each new content area, sometimes only involving a shift in interpretation, but other times entailing new statistical methods or designs. Adopting these approaches redirects research to build bridges to clinical practice and orients the practitioner to look for evidence that will change their work with the patient, thus spanning the research–practice gap from both directions.

TWELVE STEPS FOR EBM, AND A COROLLARY CLINICAL RESEARCH AGENDA

The process of teaching and using the EBA model in our clinic has augmented the steps focused on a single disorder, and no doubt there will be more facets to add in the future. A dozen themes is a good start for outlining a near-future approach to evidence based assessment in psychology. Table 1 lists the steps, a brief description of clinical action, and the corresponding clinical research agenda—reinforcing the synthesis of research and practice in this hybrid approach. Figure 2 lays out a typical sequence of working through the steps, and also maps them onto the clinical decision-making thresholds from EBM and the next clinical actions in terms of assessment and treatment. All of these steps presume that the provider has adequate training and expertise to administer, score, and interpret the assessment tools accurately, or is receiving appropriate supervision while training in their use (Krishnamurthy et al., 2004).

1. Identify the Most Common Diagnoses and Presenting Problems in Our Setting

Before concentrating on the individual client, it is important to take stock of our clinical setting. What are the common presenting problems? What are the usual diagnoses? Are there any frequent clinical issues, such as abuse, custody issues, or self injury?

After making the short list of usual suspects, then it is possible to take stock of the assessment tools and practices in the clinic. Are evidence-based assessment tools available for each of the common issues? Are they routinely used? What are the gaps in coverage, where fairly common issues could be more thoroughly and accurately evaluated? Recent work on evidence-based assessment in psychology has anthologized different instruments and reviewed the evidence for the reliability and validity of each (Hunsley & Mash, 2008; Mash & Barkley, 2007). These can help guide selection. Tests with higher reliability and validity will provide greater precision and more accurate scores for high-stakes decisions about individuals (Hummel, 1999; Kelley, 1927). Factor analyses also help explicate how different scales relate to underlying constructs and to each other, allowing for more parsimony in test selection.

Pareto's "rule of the vital few" is a helpful approximation: It is not necessary to have the resources to address every possible diagnosis or contingency, and pursuing comprehensiveness would yield sharply diminishing returns. Instead, approximately 80% of cases in most clinics will have the same ~20% of the possible clinical issues. Organizing the assessment methods to address the common diagnoses will focus limited resources to address the routine referrals and presenting problems. Making the list of typical issues more explicit also helps trainees and new clinicians to consider their work context, and it turns descriptive data into institutional wisdom that can improve the assessment process through the steps described next. Tests that do not have adequate reliability or evidence of validity cannot have utility for individual decision making. The heuristic of "is this test valid, and will it help with the patient?" (Straus et al., 2011) provides a way of identifying tests that we do not want to use, and should not continue to teach, without new evidence that shows sufficient validity. Thinking about the common presenting problems and the reliable and valid tests that assess them also would help organize a "core battery" if a clinic decides to implement a standardized intake evaluation.

Clinical research agenda. One research approach to identifying the common clinical issues is to conduct clinical epidemiological studies, looking at the rates of diagnoses and key behavioral indicators across a range of service settings. Most epidemiological research focuses on the general population, regardless of treatment status. More relevant to clinicians would be the distributions of diagnoses in outpatient practice, in special education, in residential treatment, and the other settings where we provide services.

A second research project would be to map the relatively short list of families' typical presenting concerns (Garland, Lewczyk-Boxmeyer, Gabayan, & Hawley, 2004) onto the much larger list of diagnostic possibilities. If a family comes in worried about aggression, what is the shortlist of hypotheses to consider? What are the cultural factors and beliefs about causes of behavior that change how families seek help and engage with different treatments (Carpenter-Song, 2009; Yeh et al., 2005)?

2. Know the Base Rates of the Condition in Our Setting

Meehl (1954) advocated "betting the base rate" as a simple strategy to improve the accuracy of clinical

TABLE 1
Twelve Steps in Evidence-Based Assessment and Research

Assessment Step	*Rationale*	*Clinical Research Agenda*
1. Identify most common diagnoses in our setting	Planning for the typical issues helps ensure that appropriate assessment tools are available and routinely used	Clinical epidemiology; mapping presenting problem and cultural factors onto diagnoses and research labels.
2. Know base rates	Base rate is an important starting point to anchor evaluations and prioritize order of investigation	Clinical epidemiology; meta-analyses of rates across different settings and methods.
3. Evaluate relevant risk and moderating factors	Risk factors raise "index of suspicion," enough combined elevate probability into assessment or possibly treatment zones	Compare rates of risk factors in those with versus without target diagnosis; reexpress as DLRs; meta-analyses to identify moderators.
4. Synthesize broad instruments into revised probability estimates	Already widely used; know what the scores mean in terms of changing probability for common conditions	Analyses generating DLRs for popular broad coverage instruments for different clinical targets.
5. Add narrow and incremental assessments to clarify diagnoses	Often more specific measures will show better validity, or incremental value supplementing broad measures	Test incremental validity, or superiority based on cost/benefit ratio.
6. Interpret cross-informant data patterns	Pervasiveness across settings/informants reflects greater pathology. Important to understand typical patterns of disagreement, and not overinterpret common patterns.	Test diagnostic efficiency of each informant separately; test incremental value of combinations.
7. Finalize diagnoses by adding necessary intensive assessment methods	If screening and risk factors put revised probability in the "assessment zone," what are the evidence-based methods to confirm or rule out the diagnosis in question? (e.g., KSADS, neurocognitive testing . . .)	Evaluate tests in sequence in different settings to develop optimal order and weights. Develop highly specific assessments to help rule in diagnoses.
8. Complete assessment for treatment planning and goal setting	Rule out general medical conditions, other medications; Family functioning, quality of life, personality, school adjustment, comorbidities	Develop systematic ways of screening for medical conditions and medication use. Test family functioning, personality, comorbidity, socioeconomic status and other potential moderators of treatment effects.
9. Measure processes ("dashboards, quizzes and homework")	Life charts, mood and energy checkups at each visit, medication monitoring, therapy assignments, daily report cards, three-column and five-column charts . . .	Demonstrate treatment sensitivity; meditational analyses; dismantling studies examining value added.
10. Chart progress and outcome ("midterm and final exams")	Repeat assessment with main severity measures—interview and/or parent report most sensitive to treatment effects	Jacobson and Truax (1991) benchmarks and reliable change metrics; comparison of effect sizes in same trial for different methods; develop low burden methods generalizable across patients, settings, systems. If poor response, revisit diagnoses.
11. Monitor maintenance and relapse	Discuss continued life charting; review triggers, critical events and life transitions	Event history analyses (predictors of relapse, durable recovery), key predictors, recommendations about next action if roughening.
12. Solicit and integrate patient preferences	Patient beliefs and attitudes influence treatment seeking and engagement. Possible to use these preferences to adjust wait-test and test-treat thresholds or utilities.	Qualitative analyses to identify key themes, cultural factors, preferences; studies of how to quantify preferences and add to decision making.

Note: DLR = diagnostic likelihood ratio; KSADS = Kiddie Schedule for Affective Disorders and Schizophrenia.

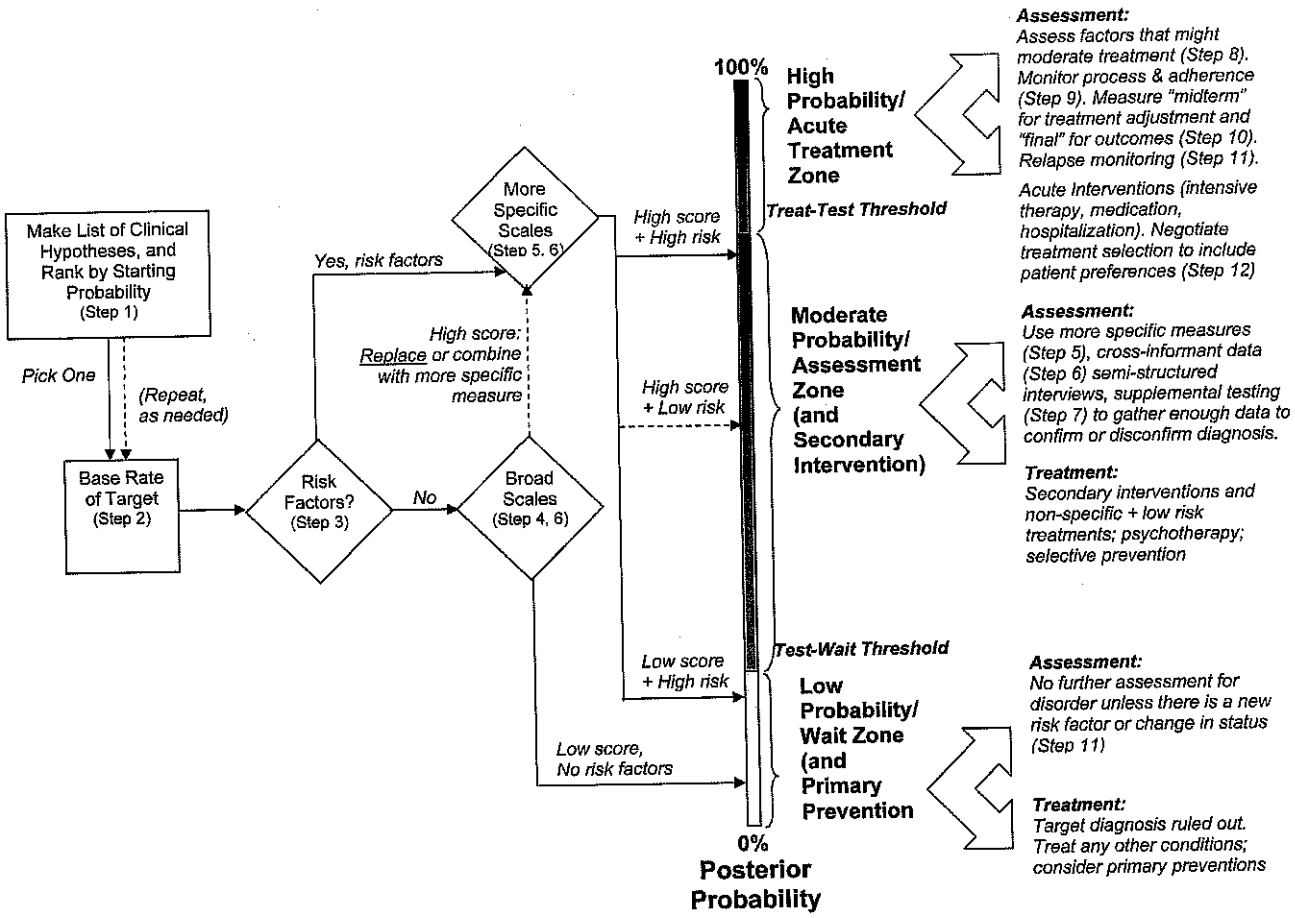

FIGURE 2 Mapping assessment results onto clinical decision making.

assessment, using the base rate as the Bayesian prior probability before adding assessment findings. When the same constellation of symptoms could be explained by an exotic or a quotidian illness, wager on the common cause. A stomachache and fever are more likely to be due to a cold virus than ebola hemhorragic fever, unless there are many other risk factors and signs that point toward the more rare explanation. The clinical epidemiological rates provide a helpful starting point for ranking the potential candidates in terms of probability before considering any case-specific information, organizing a set of potential clinical hypotheses. The prevalence of different conditions also provides a good starting estimate, taking advantage of what cognitive psychologists call the "anchoring heuristic" (Croskerry, 2003; Gigerenzer & Goldstein, 1996). Rather than interpreting case information intuitively, formally thinking about the base rates as a starting point helps increase the consistency of decision making across clinicians (Garb, 1998). Psychology has contributed both to the research about decision making and cognitive heuristics and to descriptive studies of prevalence in different settings.

Clinical research agenda. As more clinical epidemiology studies are published, then meta-analyses could describe general patterns across levels of service and identify moderating variables that change referral patterns. Studies using semistructured or structured interviews provide valuable benchmarks against which to compare local patterns. For example, if studies of urban community mental health centers find that roughly 50% of referrals meet criteria for a diagnosis of ADHD but only 20% of youths at a local center receive clinical diagnoses, or 80% for that matter, then the benchmark raises important questions about whether local assessment practices could benefit from upgrading the evidence based components.

3. Evaluate the Relevant Risk and Moderating Factors

Within the EBM framework, risk factors become data to integrate into the formal assessment process. The DLR central to the EBM method is a ratio of the diagnostic sensitivity to the false alarm rate. Put another way, the DLR compares how often the test result or risk

factor would occur in those with the diagnosis (i.e., sensitivity) versus its rate in those without the diagnosis (i.e., false alarm rate). If low birth weight was present in 3% of youths with ADHD but only 1% of those without ADHD, then the DLR attached to low birth weight would be 3.0 for ADHD. The DLR is the factor by which the odds of diagnosis change in Bayesian analysis. For clinical purposes, the conceptual status of low birth weight changes from an empirically identified "risk factor" to a variable contributing a specific weight to decision making about a particular individual case. EBM suggests that risk factors or tests producing DLRs of less than 2 are rarely worth adding to the evaluation process, whereas values around 5 are often helpful, and values greater than 10 frequently have decisive impact on an evaluation (Straus et al., 2011).

Clinical research agenda. Extensive developmental psychopathology research has focused on identifying risk and protective factors. However, these are primarily reported in terms of statistical significance and group-level effect sizes (Kraemer et al., 1999). The next step is to convert these findings into a metric amenable to idiographic assessment and decision making. The necessary statistics to generate DLRs for risk factors are simple. A chi-squared test comparing the presence or absence of the risk factor in those with or without the diagnosis is sufficient to test the validity of the risk factor (Kraemer, 1992). The next step, rarely taken in psychology to date, is to report the percentages: How common is the risk factor in those with the diagnosis versus without? Those constitute the numerator and denominator of the DLR.

4. Synthesize Broad Instruments into Revised Probability Estimates

Many clinics and practitioners use a broad assessment instrument as a standard element of their intake (e.g., Child Behavior Checklist, Behavior Assessment System for Children; Achenbach & Rescorla, 2001; Reynolds & Kamphaus, 2004). Broad instruments have a variety of strengths, including providing norm-referenced scores that compare the level of problems to what would be age- and gender-typical levels, as well as systematically assessing multiple components of emotional and behavior problems regardless of the particular referral question. This breadth prevents some cognitive heuristics that otherwise plague unstructured clinical assessments, such as concentrating only on one hypothesis, or "search satisficing" and stopping the evaluation as soon as one plausible diagnosis is identified (Croskerry, 2003; Spengler et al., 1995). The next step in an evidence-based assessment approach is to incorporate the test results and see how they raise or lower the posterior probability of the contending diagnoses. In the Bayesian EBM framework, the test score ranges have DLRs attached, and these get combined with the prior probability and risk factor DLRs to generate a revised probability estimate. It is worth noting that broad measures will not cover all possible conditions, despite their breadth. Problems that are rare in the general population may not have enough representation to generate their own "syndrome scale." This does not invalidate the use of broad measures in an EBA approach, but rather reminds us to be aware of the limits of content coverage and not unwittingly exclude clinical hypotheses outside of the scope of coverage.

Clinical research agenda. There have been a smattering of studies using Receiver Operating Characteristic (ROC) analyses to evaluate the diagnostic efficiency of broad instruments with regard to specific diagnoses such as ADHD (e.g., Chen, Faraone, Biederman, & Tsuang, 1994) and anxiety (e.g., Aschenbrand, Angelosante, & Kendall, 2005). The next step would be to calculate multilevel likelihood ratios attached to low, moderate, and high scores on the test (Guyatt & Rennie, 2002). The multilevel approach preserves more information from continuous measures, and it also is likely to be more generalizable and less sample dependent than approaches focused on picking the "optimal" cut scores (Kraemer, 1992). The approach can be simple yet still highly informative: Samples could be divided into thirds or quintiles on the Externalizing or Internalizing scale, and then the percentage of cases with the diagnosis compared to the percentage without the diagnosis in each score stratum to determine the diagnostic likelihood ratio (e.g., Youngstrom et al., 2004). As the research literature becomes more rich, then it would be possible for meta-analyses to test the generalizability of results and document moderating factors (Hasselbad & Hedges, 1995).

5. Add Narrow and Incremental Assessments to Clarify Diagnoses

At some clinics a common referral issue may not be adequately assessed by broad instruments. Pervasive developmental disorders, eating disorders, bipolar disorders, and other topics all may require the addition of more specialized measures or checklists (Mash & Hunsley, 2005). Again, a good survey of the common issues at a particular setting guides rational additions to the assessment battery. Some important issues may only be addressed by a single item or omitted entirely from broad assessment measures: The Achenbach instruments do not have scales for mania, eating

disorders, or autism, per se, for example. Psychological research has also made advances in terms of documenting incremental validity of combinations of tests (Johnston & Murray, 2003) as well as statistically testing what factors moderate the performance of tests (Cohen et al., 2003; Zumbo, 2007). The best candidates for addition to the assessment protocol will be tools that have demonstrated validity for the target diagnosis, and ideally have DLRs available so that the scores can be translated directly into a revised probability.

Clinical research agenda. Validating more narrow tests for diagnostic efficiency involves several steps. At early stages, studies performing receiver operating characteristic analyses would establish the discriminative validity of the assessment (McFall & Treat, 1999). Ideally the study design would follow the recommendations of the Standardized Reporting of Diagnostic tests guidelines (Bossuyt et al., 2003), and it would use clinically generalizable comparison groups to develop realistic estimates of performance (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006a). Later steps in the research process could include comparing the ROC performance of multiple tests either in the same sample (using procedures developed by Hanley & McNeil, 1983), or meta-analytically (Hasselbad & Hedges, 1995). Logistic regression models, using diagnosis as the dependent variable, could test whether there is incremental value in combining different tests. Logistic regression also offers a flexible framework for testing potential moderators of assessment performance, such as gender, ethnicity, culture (Garb, 1998), or credibility of the informant (Youngstrom et al., 2011). EBM teaches us to ask, "Do these results apply to this patient?" (Straus et al., 2011). The psychometric tradition has developed powerful tools to answer the question of whether results generalize, versus the validity changing due to demographic or clinical characteristics (Borsboom, 2006). When appropriate samples are available, then generating multilevel likelihood ratios for the narrow instrument also would be crucial to facilitate clinical application.

6. Interpret Cross-Informant Data Patterns

A stock recommendation in clinical assessment of youths is to gather data from multiple informants, including parents, teachers, and direct observations, as well as self-report or performance measures from the youth. However, it is well-established that these different sources of information show only modest to moderate convergence, usually in the range of $r = .1–.4$ (Achenbach, McConaughy, & Howell, 1987). Additional data can actually degrade the quality of clinical decision making, especially when the new data have low validity for the criterion of interest or when suboptimal strategies are used to synthesize information. Context and diagnostic issue moderate the validity of data across informants (De Los Reyes & Kazdin, 2005). Self-report of attention problems, or teacher report of manic symptoms, are examples of information with validity that is significantly lower than could be gleaned by asking the same questions of other sources. Adding more tests to a battery always increases the time, cost, and complexity, but it does not always improve the output (Kraemer, 1992). Cross-informant data often add considerably to the time and expense of an assessment. The psychological assessment literature has developed to a point where we can decide when the additional assessment is worth the effort, and when it would be more efficient to forego. A related point is that we can anticipate common patterns of disagreement: Whoever initiates the referral will usually be the most worried party. Low cross-informant correlations and regression to the mean will combine so that the typical scenario often looks unimpressive in terms of agreement: If the average level of parent-reported problems has a *T* score of 70, the expected level of youth or teacher reported problems would be in the range of 54 to 56 (Achenbach & Rescorla, 2001; Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006b). Recognizing and thinking through common scenarios will help avoid misinterpreting patterns in the cross-informant data (Croskerry, 2003). When different informants have shown incremental validity, then integrating the different scores into a revised probability makes sense. Even when incremental validity for diagnostic purposes may be poor, there is still value in assessing cross-informant agreement with regard to motivation for treatment (Hunsley & Meyer, 2003).

Clinical research agenda. The ideas of cross-informant data and validity are well developed in psychological assessment and virtually unknown in the traditional EBM literature. ROC and logistic regression again provide an analytic framework for evaluating the diagnostic efficiency of each informant's perspective and testing whether there is significant incremental value added by combining different informants' perspectives.

7. Finalize Diagnoses by Adding Necessary Intensive Assessment Methods

One of the goals in sequencing the assessment steps is to try to set up a "fast and frugal" order that maximizes the information value of instruments already widely used (Gigerenzer & Goldstein, 1996) and that minimizes the additional time and expense used in the first wave of assessment for a case. Based on the initial findings,

many clinical hypotheses will be "ruled out." However, few of our assessment tools are sufficiently specific to a diagnostic issue or accurate enough to confirm a diagnosis on their own. After conducting the initial evaluation, clinicians will often find that the revised probability estimate falls in the middle "assessment zone," and additional assessment is needed to confirm or disconfirm the diagnosis. More intensive and expensive tests are justified for contending diagnoses at this stage: The prior steps have screened out low probability cases so that the more expensive methods are not being used indiscriminately (Kraemer, 1992). Reserving some procedures until there are documented risk factors and suggestive findings helps establish "medical necessity" for added assessment.

One good option would be to perform a structured or semistructured diagnostic interview, or at least the modules that are relevant to the diagnostic hypotheses for the particular case at hand. Structured interviews are more reliable and valid than unstructured clinical interviews, and they do a better job of detecting comorbid diagnoses if the full version is administered (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). However, they are not a panacea: They do not have perfect validity themselves, and they can take more time than unstructured interviews (Kraemer, 1992). Also, none of them include all possible diagnoses, and any given protocol may omit at least one diagnosis that might be common at a particular setting. Until the most recent version, for example, the Kiddie Schedule for Affective Disorders and Schizophrenia (Kaufman et al., 1997) did not include a module for pervasive developmental disorders; and many interviews designed for use with youths omit bipolar disorder, eating disorders, nonsuicidal self-injury, or other conditions that have become a concern since the interviews were written or validated.

Of interest, structured approaches may be more popular with clients than with the practitioners, who cite concerns about damaging rapport as well as loss of professional autonomy as objections to routine use of more structured approaches (Suppiger et al., 2009). Structured approaches may put more administrative burden on the clinician as well as taking more time with the client (Ebesutani, Bernstein, Chorpita, & Weisz, 2012). By placing semistructured approaches at Step 7, I advocate a "combined" approach, where we consider the findings from our setting (e.g., base rates), any risk factors that might modify initial hypotheses, and the results from any checklists or rating scales *before* beginning an interview. Although Step 7 sounds late in the process, it actually falls in the first 5 to 15 min of working with an individual case. Equipped with the context and data from the prior steps, it becomes possible to decide whether to change interviews or augment with other modules or tests to cover gaps in the default interview. It also might be possible to omit modules from a semistructured interview based on revised probabilities falling below the "wait-test" threshold, although the time savings will be modest if the interview already was structured to "skip out" after a few negative responses to screening questions.

Other strategies that make sense to invoke at this stage include any other procedure that has shown incremental validity for the question of interest (Johnston & Murray, 2003) but might be too expensive or burdensome to use more generally. Essentially, this stage is a "selected or targeted" zone of assessment, analogous to selected, secondary interventions in the parlance of the International Institute of Medicine and of community mental health (Mechanic, 1989). Neurocognitive testing, daily mood charting, and soon various forms of brain imaging all might fit in this category.

Clinical research agenda. The field has been doing a good job of validating assessment strategies. The next step needed is to evaluate these tools embedded in assessment sequences tailored for distinct settings. Test consumers should not accept the developers' descriptions of test performance uncritically but rather think about how characteristics in the target and comparison group affect test performance (Bossuyt et al., 2003; Youngstrom et al., 2006a).

8. Refine Assessment for Case Formulation, Treatment Planning, and Goal Setting

There are a large number of general medical conditions and medication-related side effects that can masquerade as psychological issues. These often are measured in haphazard fashion, rather than via structured review of systems. Similarly, there are many potential treatment targets or outcome modifiers—such as personality or temperament traits, school adjustment, family functioning, parental education level—that also could be valuable to assess as part of case conceptualization and treatment selection. As we learn more about moderators of outcome, and factors that make people better matches for some treatments than others, organizing assessment to rapidly evaluate these relevant moderators will be an excellent opportunity to integrate research and practice. Assessing quality of life and functioning also is pivotal in establishing treatment goals beyond symptom reduction (Frisch, 1998).

Clinical research agenda. Much more needs to be done in terms of systematizing the evaluation of treatment moderators and also "Axis III" factors (American Psychiatric Association, 2000), such as medications and general medical conditions that have psychological

effects. Here, the initial research can move from descriptive studies to examining these variables as moderators of treatment response or predictors of optimal treatment match.

9. Measure Processes ("Quizzes, Homework, and Dashboards")

Once treatments are started, then the role of assessment changes from diagnosis to monitoring treatment progress, including mediators, process variables, and outcomes. Sometimes the intervention itself will generate products that can be used for progress checks. Examples would include behavior tracking charts, reward calendars, daily report cards, three-column and five-column charts from cognitive-behavioral therapy, and daily mood charts (Youngstrom, 2008). Many aspects of functional behavioral analysis fit well in this context, too (Vollmer & Northup, 1996). Activities completed outside of the therapy session are frequently described as "homework" to promote skill generalization. Extending the metaphor, skill assessments during sessions could be likened to "quizzes" to evaluate learning. All of these can be ratcheted toward enhancing outcome by tracking and plotting them systematically (Cone, 2001; Powsner & Tufte, 1994). Weight loss programs all measure weight repeatedly, and they have demonstrated added value of written records of food consumption and exercise on producing greater and more lasting change (Grilo, Masheb, Wilson, Gueorguieva, & White, 2011). Process measurement is much more elaborated in psychological assessment than in most of EBM, which has concentrated on diagnosis, treatment selection, and likelihood of help versus harm as the primary assessment activities (Straus et al., 2011). If the patient is failing to progress as anticipated, and especially if there are complications, we should also use this as an opportunity to reassess our case formulation and diagnoses.

Clinical research agenda. Much could be done looking at human factors that promote the uptake of some tracking methods over others. Does a smartphone application improve utilization compared to pencil and paper (e.g., Chambliss et al., 2011)? Does better utilization lead to better outcome or more durable effects? Augmentation or dismantling studies, adding or subtracting different elements of process tracking, can be embedded within other trials or routine care at clinics, helping to identify what forms of tracking are most helpful. Another promising line of work would be examining how to package these assessments into "dashboards" that provide a clear summary of progress easily interpreted by family and therapist alike (Few, 2006; Powsner & Tufte, 1994).

10. Chart Progress and Outcome ("Midterm and Final Exams")

Continuing with the education metaphor, outcome evaluation can be cast as the "final exam," measuring the amount of change over the course of treatment. There are several operational definitions of outcome, including loss of diagnosis, percentage reduction of symptoms on a severity measure, or more complex definitions of "clinically significant change" that combine information about the precision of the measure—such as the "reliable change index"—with comparisons to normative benchmarks based on distributions in clinical and nonclinical samples (Jacobson & Truax, 1991). All of these involve more lengthy and comprehensive evaluation than the "process" measures just described, and so these panels of assessment methods are used more episodically. In clinical practice, outcome evaluation is more likely to be informal, based on the view that it is obvious when people are improving, and the belief that clients and payers will not accept the additional assessment involved (Suppiger et al., 2009). Contrary to expectation, clients are likely to view thorough assessments positively (Suppiger et al., 2009), and payers are more likely to reimburse assessments that are clearly linked to treatment (Cashel, 2002). Services databases consistently show modest rates of improvement and great heterogeneity in outcomes for treatment as usual, with some cases improving markedly, and others actually deteriorating. Meehl and others have argued that the slow progress in psychological treatment is due in large part to our failure to measure outcomes and get corrective feedback about when our interventions help, are inert, or even harm (Christensen & Jacobson, 1994; Meehl, 1973).

Research about patterns of treatment response also indicates potential value in having a scheduled "midterm," where more intensive evaluation is done to quantify early response to treatment. Early response to intervention, both psychotherapy and pharmacological (Curry et al., 2011), often predicts long-term response (Howard, Moras, Brill, Martinovich, & Lutz, 1996). If a person does not show improvement over the first 4 to 8 weeks or sessions, then it makes sense to either augment or change the modality of treatment (Lambert, Hansen, & Finch, 2001). Careful assessment of early response is also crucial to monitoring side effects and potential treatment-emergent changes in mood or behavior that should trigger alterations in the treatment plan (Joseph, Youngstrom, & Soares, 2009). Outcome evaluation is another area where psychological assessment has developed more sophisticated models for evaluating individual change compared to the metrics commonly used in EBM. Number needed to treat (the number of people who would need exposure to the treatment for

one more case to have a good outcome), number needed to harm (the number of people who would need exposure to the treatment for one more case to experience harmful side effects or iatrogenic outcomes), and similar indices are all measures of probabilistic efficacy based on groups of cases and dichotomous outcomes (Guyatt & Rennie, 2002). Psychological assessment offers much in terms of benchmarking against typical ranges of functioning, looking at change on continuous measures, and considering the precision of measurement when evaluating individual outcomes.

Clinical research agenda. There are a variety of methods worth investigating, including trials examining whether the addition of assessment at the "midterm" or end of acute treatment changes engagement, adherence, and acute or long-term outcomes (e.g., Ogles, Melendez, Davis, & Lunnen, 2001). A second line of work could optimize instruments for outcome evaluation by demonstrating sensitivity to treatment effects, developing shorter versions that retain sufficient precision to guide individual treatment decisions, and establishing meaningful benchmarks for "clinically significant change" approaches.

11. Monitor Maintenance and Relapses

Many disorders of childhood and adolescence carry a high risk of relapse, such as mood disorders; others are associated with an elevated risk of developing later pathology, perhaps as forms of heterotypic continuity. Anxiety often augurs later depression (Mineka, Watson, & Clark, 1998), and ADHD often presages substance issues or conduct problems (Taurines et al., 2010). More could be done in terms of educating families around signs of relapse or cues of early onset of later problems. Creative work is being done with mood disorders, helping patients identify signs of "roughening" and changes in energy or behavior that might offer early warning of relapse (Sachs, 2004), and then planning ahead of time for strategies that can help restabilize mood or promote earlier intervention to minimize the effects of recurrence. Given what we know about the epidemiology of mental health problems and developmental changes through adolescence and early adulthood, a combination of general screening and brief, targeted evaluations of warning signs could accomplish much good. This aspect of assessment has not received much attention from either the EBM or psychological assessment traditions yet, and represents a major growth area.

Clinical research agenda. It would be intriguing to evaluate how customized assessment strategies might predict shorter lag to seeking treatment, increased utilization of prevention or early intervention services, or diversion from more acute and tertiary treatments. Similarly, it would be important to know whether brief, broad coverage measures might have a role in primary care or other settings as predictors of relapse or progression in youths who have previously benefitted from treatment. Advances in technology make a variety of "smart" applications feasible as methods for monitoring behavior for cues of relapse.

12. Solicit and Integrate Patient Preferences

The placement of the wait-test and treat-test thresholds is flexible in EBM (Straus et al., 2011) (see also Figure 2). Their location is supposed to be guided by the costs and benefits attached to the diagnosis or treatment, as well as patient preferences. For dichotomous outcomes, such as recovery or remission, there is a developed framework combining the number needed to treat with the number needed to harm, yielding a Likelihood of Help versus Harm that can be further adjusted based on patient preferences (Straus et al., 2011). There are other formal mathematical approaches to synthesizing costs, benefits, and assessment parameters to optimize decision thresholds (Kraemer, 1992; Swets, Dawes, & Monahan, 2000), too. The EBM approach is attractive because it is simple enough that it could be done in session with families, potentially working through several "what if..." scenarios together to help explore a range of options and guide consensual decisions.

There is a rich layer of additional information that could be added here, using surveys and interviews to solicit beliefs about causes of emotional and behavioral problems, differences in what is perceived as problematic, and attitudes toward help-seeking and different services. Beliefs about medication and therapy have great influence over treatment seeking and engagement (Yeh et al., 2005). The effects of culture on decisions to seek or continue treatment are likely to be as big or bigger than culture's moderating effects on the accuracy of assessments or intervention efficacy. This aspect of assessment is one of the most promising places to combine psychological assessment's sophistication about measuring beliefs, attitudes, and preferences with the mathematical framework and decision aids offered by EBM.

Clinical research agenda. Qualitative methods as well as quantitative interviews and surveys have much to add in terms of knowledge about patient preferences. There also is a great deal that could be done integrating preferences into the decision-making framework, adjusting the test score thresholds for screening programs at a policy level (Swets et al., 2000) or negotiating personalized decision making with individual cases (Straus et al.,

2011). The algorithms have been available for decades, but it is only recently that technology has made it convenient for families and practitioners to use the tools. Recent developments understanding the role of culture in service selection, stigma, and attitudes to treatment also provides more rich inputs into the decision-making process (Hinshaw & Cicchetti, 2000; Yeh et al., 2005). Although last in the "steps" listed here, understanding patient attitudes is something we could profitably weave through the entire assessment process.

DISCUSSION

When it convened more than a dozen years ago, the Psychological Assessment Work Group of the American Psychological Association concluded there was surprisingly little published data to document the value of conventional psychological assessment in terms of better outcomes (Eisman et al., 1998; Meyer et al., 1998). The situation has improved only modestly in subsequent years (Hunsley & Mash, 2007). Our failure to measure things that matter to families and for treatment still contributes to the slow progress of our interventions (Meehl, 1973; Nelson-Gray, 2003).

EBM lacks the psychometric sophistication that has characterized the best traditions of psychological assessment. Psychological assessment has developed a wide range of instruments, and psychometric models could provide sophisticated techniques for honing the analytical underpinnings of EBM (Borsboom, 2008). What EBM offers, though, is a pragmatic focus on understanding and helping the individual case. EBM ties assessment to clinical decision making with a directness and clarity that has been missing in much of psychological assessment. Integration is possible, keeping the psychometric and conceptual strengths of psychological assessment but incorporating them into the decision-making framework articulated in EBM. The fit is not seamless, but it is patient centered, clinically relevant, and compelling. Some of the looser connections will be promising areas of investigation in their own right. EBM has historically emphasized dichotomous outcomes (e.g., recovery, death), whereas psychology has focused more on continuous measures. It is possible to convert dimensional effect sizes, such as Cohen's *d* or a correlation coefficient, into other effect sizes such as risk ratios (Hasselbad & Hedges, 1995), making it possible to reexpress outcomes in metrics that fit within the EBM decision-making framework, but it also would be intriguing to develop parallel approaches that capitalize on the greater information intrinsic to continuous measures.

Exploring the potential for synthesis reorganized my approach to assessment research, teaching, and supervision. Viewing assessment through an EBM tinted lens defines a set of clinical research topics that comprise a thematic program of investigation. The research designs and statistical methods are readily available and not complex. Adopting these methods need not add to the expense of the assessment process: Better decisions can be made by using the same tools but interpreting them differently. For example, we have found that there can be pronounced changes in clinical decisions about vignettes, with increased accuracy and consistency, and an elimination of a tendency to overdiagnose bipolar disorder, based on identical assessment data combined with brief training in the probability nomogram as a way of interpreting scores (Jenkins et al., 2011). The value of these methods is not limited to bipolar disorder, any more than it would be limited to any single area within medicine (Guyatt & Rennie, 2002). The hybridization of psychological assessment with EBM ideas produces ideas with vigor and clinical relevance to rejuvenate assessment and ultimately improve outcomes for families (Bauer, 2007).

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington: University of Vermont.

Algorta, G. P., Youngstrom, E. A., Phelps, J., Jenkins, M. M., Youngstrom, J. K., & Findling, R. L. (2012). An inexpensive family index of risk for mood issues improves identification of pediatric bipolar disorder. *Psychological Assessment*. Advance online publication. doi:10.1037/a0029225

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychological Association. (2005). Policy statement on evidence-based practice in psychology. Retrieved from http://www.apa.org/practice/resources/evidence/evidence-based-statement.pdf

Aschenbrand, S. G., Angelosante, A. G., & Kendall, P. C. (2005). Discriminant validity and clinical utility of the CBCL with anxiety-disordered youth. *Journal of Clinical Child and Adolescent Psychology, 34*, 735–746. doi:10.1207/s15374424jccp3404_15

Bauer, R. M. (2007). Evidence-based practice in psychology: implications for research and research training. *Journal of Clinical Psychology, 63*, 685–694. doi:10.1002/jclp.20374

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London, 53*, 370–418. doi:10.1098/rstl.1763.0053

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57*, 717–726.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*, 1089–1108. doi:10.1002/jclp.20503

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., ... de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41–44. doi:10.1136/bmj.326.7379.41

Camara, W., Nathan, J., & Puente, A. (1998). *Psychological test usage in professional psychology: Report of the APA practice and science directorates* (p. 51). Washington, DC: American Psychological Association.

Carpenter-Song, E. (2009). Caught in the psychiatric net: meanings and experiences of ADHD, pediatric bipolar disorder and mental health treatment among a diverse group of families in the United States. *Culture, Medicine and Psychiatry, 33*, 61–85. doi:10.1007/s11013-008-9120-4

Cashel, M. L. (2002). Child and adolescent psychological assessment: Current clinical practices and the impact of managed care. *Professional Psychology: Research and Practice, 33*, 446–453. doi:10.1037//0735-7028.33.5.446

Chambliss, H. O., Huber, R. C., Finley, C. E., McDoniel, S. O., Kitzman-Ulrich, H., & Wilkinson, W. J. (2011). Computerized self-monitoring and technology-assisted feedback for weight loss with and without an enhanced behavioral component. *Patient Education and Counseling, 85*, 375–382. doi:10.1016/j.pec.2010.12.024

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology, 62*, 1017–1025. doi:10.1037/0022-006X.62.5.1017

Childs, R. A., & Eyde, L. D. (2002). Assessment training in clinical psychology doctoral programs: what should we teach? What do we teach? *Journal of Personality Assessment, 78*, 130–144. doi:10.1207/S15327752JPA7801_08

Christensen, A., & Jacobson, N. S. (1994). Who (or what) can do psychotherapy: The status and challenge of nonprofessional therapies. *Psychological Science, 5*, 8–14. doi:10.1111/j.1467-9280.1994.tb00606.x

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cone, J. D. (2001). *Evaluating outcomes: Empirical tools for effective practice*. Washington, DC: American Psychological Association.

Correll, C. U. (2008). Antipsychotic use in children and adolescents: Minimizing adverse effects to maximize outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*, 9–20. doi:10.1097/chi.0b013e31815b5cb1

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*, 775–780. doi:10.1097/00001888-200308000-00003

Curry, J., Silva, S., Rohde, P., Ginsburg, G., Kratochvil, C., Simons, A., ... March, J. (2011). Recovery and recurrence following treatment for adolescent major depression. *Archives of General Psychiatry, 68*, 263–269. doi:10.1001/archgenpsychiatry.2010.150

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674. doi:10.1126/science.2648573

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131*, 483–509. doi:10.1037/0033-2909.131.4.483

Ebesutani, C., Bernstein, A., Chorpita, B. F., & Weisz, J. R. (2012). A transportable assessment protocol for prescribing youth psychosocial treatments in real-world settings: Reducing assessment burden via self-report scales. *Psychological Assessment, 24*, 141–155. doi:10.1037/a0025176

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., ... Moreland, K. L. (1998). *Problems and limitations in the use of psychological assessment in contemporary health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part II* (p. 22). Washington, DC: American Psychological Association.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. Cambridge, MA: O'Reilly Press.

Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. R. (2005). Evidence-based assessment of learning disabilities in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*, 506–522. doi:10.1207/s15374424jccp3403_7

Frazier, T. W., & Youngstrom, E. A. (2006). Evidence-based assessment of attention-deficit/hyperactivity disorder: Using multiple sources of information. *Journal of the American Academy of Child & Adolescent Psychiatry, 45*, 614–620. doi:10.1097/01.chi.0000196597.09103.25

Frisch, M. B. (1998). Quality of life therapy and assessment in health care. *Clinical Psychology: Science and Practice, 5*, 19–40.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment*. Washington, DC: American Psychological Association.

Garland, A. F., Lewczyk-Boxmeyer, C. M., Gabayan, E. N., & Hawley, K. M. (2004). Multiple stakeholder agreement on desired outcomes for adolescents' mental health services. *Psychiatric Services, 55*, 671–676. doi:10.1176/appi.ps.55.6.671

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669. doi:10.1037/0033-295X.103.4.650

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. (1997). Incremental efficacy of WISC–III factor scores in predicting achievement: What do they tell us? *Psychological Assessment, 9*, 295–301. doi:10.1037/1040-3590.9.3.295

Gray, G. E. (2004). *Evidence-based psychiatry*. Washington, DC: American Psychiatric Publishing.

Grilo, C. M., Masheb, R. M., Wilson, G. T., Gueorguieva, R., & White, M. A. (2011). Cognitive-behavioral therapy, behavioral weight loss, and sequential treatment for obese patients with binge-eating disorder: A randomized controlled trial. *Journal of Consulting & Clinical Psychology, 79*, 675–685. doi:10.1037/a0025049

Guyatt, G. H., & Rennie, D. (Eds.). (2002). *Users' guides to the medical literature*. Chicago, IL: AMA Press.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Harkness, A. R., & Lilienfeld, S. O. (1997). Individual differences science for treatment planning: Personality traits. *Psychological Assessment, 9*, 349–360.

Hasselbad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167–178. doi:10.1037/0033-2909.117.1.167

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42*, 963–974.

Hinshaw, S. P., & Cicchetti, D. (2000). Stigma and mental disorder: Conceptions of illness, public attitudes, personal disclosure, and social policy. *Development & Psychopathology, 12*, 555–598. doi:10.1017/S0954579400004028

Hodgins, S., Faucher, B., Zarac, A., & Ellenbogen, M. (2002). Children of parents with bipolar disorder. A population at high risk for major affective disorders. *Child & Adolescent Psychiatric Clinics of North America, 11*, 533–553.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and

patient progress. *American Psychologist, 51*, 1059–1064. doi:10.1037/0003-066X.51.10.1059

Hummel, T. J. (1999). The usefulness of tests in clinical decisions. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston, MA: Allyn and Bacon.

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment, 17*, 251–255. doi:10.1037/1040-3590.17.3.251

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psychology, 3*, 29–51. doi:10.1146/annurev.clinpsy.3.022806.091419

Hunsley, J., & Mash, E. J. (Eds.). (2008). *A guide to assessments that work*. New York, NY: Oxford University Press.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19. doi:10.1037/0022-006X.59.1.12

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature: III. How to use an article about a diagnostic test: B: What are the results and will they help me in caring for my patients? *Journal of the American Medical Association, 271*, 703–707.

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42*, 121–129. doi:10.1037/a0022506

Jenkins, M. M., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2012). Generalizability of evidence-based assessment recommendations for pediatric bipolar disorder. *Psychological Assessment, 24*, 269–281. doi:10.1037/a0025775

Johnston, C., & Murray, C. (2003). Incremental validity in the psychological assessment of children and adolescents. *Psychological Assessment, 15*, 496–507.

Joseph, M., Youngstrom, E. A., & Soares, J. C. (2009). Antidepressant-coincident mania in children and adolescents treated with selective serotonin reuptake inhibitors. *Future Neurology, 4*, 87–102. doi:10.2217/14796708.4.1.87

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., ... Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*, 980–988. doi:10.1097/00004583-199707000-00021

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Books.

Kovacs, M. (1992). *Children's Depression Inventory Manual*. North Tonawanda, NY: Multi-Health Systems.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods, 4*, 257–271.

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., ... Benton, S. A. (2004). Achieving competency in psychological assessment: directions for education and training. *Journal of Clinical Psychology, 60*, 725–739. doi:10.1002/jclp.20010

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: using patient outcome data to enhance treatment effects. *Journal of Consulting & Clinical Psychology, 69*, 159–172. doi:10.1037/0022-006X.69.2.159

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614. doi:10.1146/annurev.psych.58.110405.085542

Mash, E. J., & Barkley, R. A. (Eds.). (2007). *Assessment in children and adolescents*. New York, NY: Guilford.

Mash, E. J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psychology, 34*, 362–379. doi:10.1207/s15374424jccp3403_1

McFall, R. (1991). Manifesto for a science of clinical psychology. *The Clinical Psychologist, 44*, 75–88.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology, 50*, 215–241. doi:10.1146/annurev.psych.50.1.215

Mechanic, D. (1989). *Mental health and social policy*. Englewood Cliffs, NJ: Prentice-Hall.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Meehl, P. (1973). Why I do not attend case conferences. In P. Meehl (Ed.), *Psychodiagnosis: Selected papers* (pp. 225–302). New York, NY: Norton.

Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4*, 91–98. doi:10.1111/j.1468-2850.1997.tb00103.x

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 55*, 194–216.

Merenda, P. F. (2007a). Psychometrics and psychometricians in the 20th and 21st centuries: How it was in the 20th century and how it is now. *Perceptual & Motor Skills, 104*, 3–20. doi:10.2466/pms.104.1.3-20

Merenda, P. F. (2007b). Update on the decline in the education and training in psychological measurement and assessment. *Psychological Reports, 101*, 153–155. doi:10.2466/pr0.101.1.153-155

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., ... Dies, R. R. (1998). *Benefits and costs of psychological assessment in health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part I* (p. 90). Washington, DC: American Psychological Association.

Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 69*, 1–38. doi:10.1207/s15327752jpa6901_1

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*, 377–412. doi:10.1146/annurev.psych.49.1.377

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15*, 521–531.

Ogles, B. M., Melendez, G., Davis, D. C., & Lunnen, K. M. (2001). The Ohio Scales: Practical outcome assessment. *Journal of Child & Family Studies, 10*, 199–212.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Piotrowski, C. (1999). Assessment practices in the era of managed care: Current status and future directions. *Journal of Clinical Psychology, 55*, 787–796.

Powsner, S. M., & Tufte, E. R. (1994). Graphical summary of patient status. *The Lancet, 344*, 368–389. doi:10.1016/S0140-6736(94)91406-0

Ravens-Sieberer, U., & Bullinger, M. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: First psychometric and content analytic results. *Quality of Life Research, 7*, 399–407.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*, 169–184. doi:10.1002/mpr.289

Reynolds, C. R., & Kamphaus, R. (2004). *BASC–2 Behavior Assessment System for Children.* Circle Pines, MN: American Guidance Service.

Sachs, G. S. (2004). Strategies for improving treatment of bipolar disorder: Integration of measurement and management. *Acta Psychiatrica Scandinavica*, 7–17. doi:10.1111/j.1600-0447.2004.00409.x

Sattler, J. M. (2002). *Assessment of children: Behavioral and Clinical Applications* (4th ed.). La Mesa, CA: Publisher Inc.

Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A. (1995). A scientist–practitioner model of psychological assessment: Implications for training, practice and research. *The Counseling Psychologist, 23*, 506–534. doi:10.1177/0011000095233009

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; What you need to know. *Journal of Clinical Psychology, 63*, 611–631.

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment, 77*, 398–407. doi:10.1207/S15327752JPA7703_02

Stedman, J. M., Hatch, J. P., Schoenfeld, L. S., & Keilin, W. G. (2005). The structure of internship training: Current patterns and implications for the future of clinical and counseling psychologists. *Professional Psychology: Research and Practice, 36*, 3–8. doi:10.1037/0735-7028.36.1.3

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York, NY: Churchill Livingstone.

Suppiger, A., In-Albon, T., Hendriksen, S., Hermann, E., Margraf, J., & Schneider, S. (2009). Acceptance of structured diagnostic interviews for mental disorders in clinical practice and research settings. *Behavior Therapy, 40*, 272–279. doi:S0005-7894(08)00088-9 [pii]

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26. doi:10.1111/1529-1006.001

Taurines, R., Schmitt, J., Renner, T., Conner, A. C., Warnke, A., & Romanos, M. (2010). Developmental comorbidity in attention-deficit/hyperactivity disorder. *Attention Deficit and Hyperactivity Disorders, 2*, 267–289. doi:10.1007/s12402-010-0040-0

Vollmer, T. R., & Northup, J. (1996). Some implications of functional analysis for school psychology. *School Psychology Quarterly, 11*, 76–92.

Wagner, K. D., Hirschfeld, R., Findling, R. L., Emslie, G. J., Gracious, B., & Reed, M. (2006). Validation of the mood disorder questionnaire for bipolar disorders in adolescents. *Journal of Clinical Psychiatry, 67*, 827–830. doi:10.4088/JCP.v67n0518

Watkins, M. W. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15*, 465–479. doi:10.1037/h0088802

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The comprehensive system for the Rorschach: A critical examination. *Psychological Science, 7*, 3–10. doi:10.1111/j.1467-9280.1996.tb00658.x

Yeh, M., Hough, R. L., Fakhry, F., McCabe, K. M., Lau, A. S., & Garland, A. F. (2005). Why bother with beliefs? Examining relationships between race/ethnicity, parental beliefs about causes of child problems, and mental health service use. *Journal Consulting and Clinical Psychology, 73*, 800–807. doi:10.1037/0022-006X.73.5.800

Youngstrom, E. A. (2007). Pediatric bipolar disorder. In E. J. Mash & R. A. Barkley (Eds.), *Assessment of childhood disorders* (4th ed., pp. 253–304). New York, NY: Guilford.

Youngstrom, E. A. (2008). Evidence-based strategies for the assessment of developmental psychopathology: Measuring prediction, prescription, and process. In D. J. Miklowitz, W. E. Craighead, & L. Craighead (Eds.), *Developmental psychopathology* (pp. 34–77). New York, NY: Wiley.

Youngstrom, E. A., & Duax, J. (2005). Evidence based assessment of pediatric bipolar disorder, Part 1: Base rate and family history. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 712–717. doi:10.1097/01.chi.0000162581.87710.bd

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., & Price, M. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*, 847–858. doi:10.1097/01.chi.0000125091.35109.1e

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R. (2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment, 13*, 267–276.

Youngstrom, E. A., Freeman, A. J., & Jenkins, M. M. (2009). The assessment of children and adolescents with bipolar disorder. *Child and Adolescent Psychiatric Clinics of North America, 18*, 353–390. doi:10.1016/j.chc.2008.12.002

Youngstrom, E. A., Jenkins, M. M., Jensen-Doss, A., & Youngstrom, J. K. (2012). Evidence-based assessment strategies for pediatric bipolar disorder. *Israel Journal of Psychiatry & Related Sciences, 49*, 15–27.

Youngstrom, E. A., & Kogos Youngstrom, J. (2005). Evidence-based assessment of pediatric bipolar disorder, Part 2: Incorporating information from behavior checklists. *Journal of the American Academy of Child & Adolescent Psychiatry, 44*, 823–828. doi:10.1097/01.chi.0000164589.10200.a4

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006a). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry, 60*, 1013–1019. doi:10.1016/j.biopsych.2006.06.023

Youngstrom, E. A., Meyers, O., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006b). Diagnostic and measurement issues in the assessment of pediatric bipolar disorder: Implications for understanding mood disorder across the life cycle. *Development and Psychopathology, 18*, 989–1021. doi:10.1017/S0954579406060494

Youngstrom, E. A., Youngstrom, J. K., Freeman, A. J., De Los Reyes, A., Feeny, N. C., & Findling, R. L. (2011). Informants are not all equal: predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology, 21*, 407–415. doi:10.1089/cap.2011.0032

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223–233.

APPENDIX

Case Example

Referral Question: Tandi is a 10-year-old girl living with her biological parents and older sister who is coming for an outpatient evaluation because her mother is concerned about her increasing "mood swings." Tandi is in

regular education at a public school, taking accelerated classes. Her mother describes her as having been outgoing and cheerful as a child, but recently seems to have become more quiet, irritable, and crabby, sometimes snapping at her family, and recently slamming doors and throwing things. According to her mom, the paternal aunt has been diagnosed with bipolar disorder, and her mom has heard that this runs in families. She wants to know if Tandi has bipolar disorder.

Steps 1 & 2. Identify the Most Common Diagnoses and Presenting Problems in Our Setting, and Know the Base Rates of the Condition in Our Setting. The clinic where Tandi's family presented uses an electronic medical record, so it is possible to produce a report listing the most frequent diagnoses. The most common diagnosis is adjustment disorder (~60% of cases), followed by attention deficit/hyperactivity disorder (ADHD; 40%), oppositional defiant disorder (ODD; 35%), and major depressive disorder (30%, but lower in younger children and higher postpubertally). Posttraumatic stress disorder (PTSD), conduct disorder, and bipolar spectrum disorders are all diagnosed in roughly 10% of cases. The clinician has compared these rates with published rates from other outpatient settings and knows that the rank order seems plausible compared to external benchmarks. The somewhat higher rates of externalizing problems and lower rates of anxiety disorders reflect logical patterns in local referral sources. Based on this, bipolar disorder is worth assessing to address the referral question, but it is not a leading candidate. The clinic has stocked rating scales and assessment tools for all of the diagnoses that occur in 10% or more of cases, so the resources are available to explore bipolar disorder further if warranted.

Step 3. Evaluate the Relevant Risk and Moderating Factors *(Also Illustrating the Use of the Probability Nomogram).* Family history of bipolar disorder is a well-established risk factor, based on decades of research and multiple reviews. A clear diagnosis of bipolar in a first degree relative is associated with a diagnostic likelihood ratio (DLR) of 5.0, indicating a fivefold increase in the odds of the youth having a bipolar disorder (Youngstrom & Duax, 2005). A second-degree relative, such as the paternal aunt, will share on average half as many genes with the person being assessed, and thus confer half as much risk. The clinician asks the mother for more details about the aunt. Per mother's report, the aunt has been psychiatrically hospitalized twice and treated with lithium as well as an atypical antipsychotic—all details that support a bipolar diagnosis. Conceptually, the aunt's history is a "yellow flag" increasing the index of suspicion for bipolar disorder. The clinician asks the mother to complete the half-page Family Index of Risk for Mood (Algorta et al., 2012) as a way of gathering information about other relatives. The aunt is the closest relative clearly affected by mood disorder, although other relatives have histories of substance use or depression. The clinician uses the probability nomogram (Figure 1) to estimate how the family history changes the probability of a bipolar disorder for Tandi. The clinician begins by plotting the base rate of bipolar spectrum disorder at the clinic on the left hand line of the nomogram, placing a dot at the 10%. The aunt's history of bipolar disorder would have a DLR of 2.5 (or half of the 5.0 attached to a first degree relative having bipolar disorder). The 2.5 is plotted on the middle line of the nomogram. Connecting the dots and extending across the right hand line yields an estimate of ~24% for the new, "posterior" probability of bipolar disorder. If the clinician used an online calculator instead of the nomogram, then he or she would generate a probability of 22%, not very different. The FIRM score of 8 also has a DLR of 2.5; plugging that DLR into the nomogram would lead to a probability of ~22 to 24%. Note that the clinician does not treat the FIRM score and the aunt's diagnosis as separate pieces of information. Instead, the clinician either chooses to focus on the one that seems more valid or uses each separately to generate two probabilities that "bracket" Tandi's risk in a form of sensitivity analysis that examines how sensitive the estimates are to changes in the inputs. Here, both results are close together. Both also are above the clinician's wait-test threshold. More assessment is needed to decide whether bipolar is present or absent for Tandi.

Family history of bipolar disorder also increases the risk of depression, ADHD, and a variety of other conditions, typically with a DLR in the range of 1.5 to 3.0 based on a meta-analysis (Hodgins, Faucher, Zarac, & Ellenbogen, 2002). However, because it is Tandi's aunt, not a first-degree relative, the conferred risk would be half as high (falling in the 1.25 to 1.5 range). This is low enough that the clinician decides to concentrate on looking for more valid information rather than spending time combining these DLRs with the prior probabilities for the other diagnoses (Straus et al., 2011).

Step 4. Synthesize Broad Instruments into Revised Probability Estimates. Tandi's mother completed the Child Behavior Checklist (CBCL) as part of the core intake battery the clinic uses. The *T* scores are 63 for Externalizing, 67 for Internalizing, 70 for Anxious/Depressed, 67 for Withdrawn/Depressed, 51 for Attention Problems, 66 for Aggressive Behavior, and 53 for Rule Breaking. Impressionistically, the scores could be consistent with an adjustment disorder (which is still the leading hypothesis) or depression. The Externalizing scores look mild for ODD, and the low Attention Problems decreases suspicion of ADHD substantially. The low Rule Breaking

score also decreases the probability of conduct disorder, which already was uncommon at the clinic (base rate of 10%). The clinician considers conduct disorder "ruled out" unless there is new information that increases concern about it. Adjustment disorder, depression, ODD, ADHD, and bipolar are still the focus of assessment. The clinician does a PubMed search on "Child Behavior Checklist" AND "bipolar disorder" AND "sensitivity and specificity" and finds a paper that published DLRs for the CBCL Externalizing score compared to a semi-structured diagnostic criterion (Youngstrom et al., 2004). The *T* of 63 is actually in the low range for youths with bipolar disorder, and it is more than twice as likely for youth to score in this range if they do not have a bipolar diagnosis (DLR = 0.47). The clinician uses the probability of 24% (from Step 3) as the new starting point on the nomogram left hand line, and plots the DLR of 0.47 on the midline, producing a revised estimate of ~15%. If the clinician used a calculator instead for all of the steps, the probability estimate would be 12%. Using similar approaches, the clinician finds that the probability of depression is up to about 65%, ADHD is down to below 20%, and no information is readily available for predicting adjustment disorder with the CBCL. To this point, the clinician has neither added any extra assessment tools to the battery except the FIRM nor spent any additional time interviewing the family. The steps have made the list of hypotheses and the interpretation more systematic than would otherwise often be the case, and relying on base rates and published weights counteracts potential cognitive heuristics due to the family's description of the presenting problem.

Step 5. Add Narrow and Incremental Assessments to Clarify Diagnoses. Based on the current hypotheses and probability estimates, the clinician decides to add some mood rating scales evaluating both depressive and hypomanic/manic symptoms as well as gather a teacher report about Tandi's school functioning. The clinician opts for the Achenbach Teacher Report Form as a concise way of gathering data about attention problems (potentially ruling ADHD out if low, vs. indicating continued assessment if high) as well as the degree of pervasiveness of the aggressive behaviors (helpful for the ODD hypothesis). The literature suggests that the teacher report of mood symptoms is unlikely to be helpful for differential diagnosis but could be helpful for treatment planning.

The clinician has Tandi complete the Child Depression Inventory (CDI; Kovacs, 1992) and the Mood Disorder Questionnaire (MDQ; Wagner et al., 2006), which has the easiest reading level of the hypomania/mania rating scales having published data with youths (Youngstrom, 2007). The clinician asks the mom to complete the Parent General Behavior Inventory, which asks about both depressive and hypomanic symptoms (PGBI; Youngstrom, Findling, Danielson, & Calabrese, 2001). Because the mother is specifically concerned about the possibility of bipolar disorder, the clinician and mother agree to have her do the full-length version rather than one of the abbreviated ones, to provide the most comprehensive description even though there is no statistical advantage of the longer versus shorter versions. Mom's scores for Tandi on the PGBI are 16 on the Hypomanic/Biphasic Scale (28 items) and 39 on the Depression Scale (46 items). The Hypomanic/Biphasic score falls in the low range for bipolar disorder, with a DLR of .46. Using the nomogram, this reduces the probability of a bipolar disorder to ~ 7%. Tandi's scores come back moderately high on the CDI and below threshold on the MDQ. Using the sensitivity (38%) and specificity (74%) published by Wagner et al. (2006) yields a DLR of 0.84. This is close enough to 1.0 that the clinician could ignore it rather than feeding it into the nomogram or a calculator; impressionistically, it is revising the low probability of bipolar disorder to become slightly lower still. The scores on the CDI and PGBI Depression are both suggestive of depression, raising the probability to ~85%.

Step 6. Interpret Cross-Informant Data Patterns. The Teacher Report Form (TRF) comes back with all scores below a *T* of 60. Tandi's grades have been good (all 3 s and 4 s on a 4-point scale). The low score on Attention Problems from the teacher, combined with the other assessment data, reduces the probability of ADHD below 5%. The clinician considers it functionally ruled out, based on the probability and the absence of any "red flags" in the academic record. The low scores do not change the probability of a mood disorder. They slightly reduce the chances of ODD. Tandi's high self-report of depressive symptoms is consistent with her mom's report of internalizing concerns, suggesting that Tandi may be motivated for treatment working on internalizing issues.

Step 7. Finalize Diagnoses by Adding Necessary Intensive Assessment Methods. The clinician selects the depression module of the MINI as a brief, structured interview to formally cover the diagnostic criteria for major depression and dysthymic disorder, along with the ODD module. The clinician also asks about recent life events and potential stressors, looking for possible precipitants for an adjustment disorder. At this stage, the clinician also considers other rival hypotheses that could be consistent with the presentation. Before diagnosing depression, we are supposed to rule out the possibility of medication side effects or general medical conditions. The clinician explains the rationale for doing the interview and asks about medications, vitamins, or other

drugs that Tandi might be taking. Tandi has had regular pediatrician visits, and her health has been good. She is not taking any prescription medication, and to her mom's knowledge, neither her peer group nor her older sister's is using any illicit substances. The MINI results identify a sufficient number of symptoms and duration for a diagnosis of a major depressive episode, with impairment at home. The severity appears mild to moderate based on the rating scales as well as descriptions during the MINI and the clinician's observations of Tandi. Based on assessment findings, the clinician assigns a diagnosis of major depressive disorder, single episode, moderate severity. The ODD module does not pass threshold, and the clinician formulates the irritability as being a feature of the depression rather than a separate diagnostic issue.

Step 8. Refine Assessment for Treatment Planning and Goal Setting. Based on the information so far, depression seems to be a main concern. The CDI and CBCL Internalizing provide good baseline scores for severity of the problem. The clinician has charts indicating the number of points each measure needs to change to demonstrate improvement (Youngstrom, 2007), based on the reliable change index approach, as well as benchmarks for treatment targets for "clinically significant change" on those as primary outcome measures (Jacobson & Truax, 1991). The clinician supplements this with measures of quality of life to look at positive aspects of functioning (Frisch, 1998) and selects the KINDL as a brief, developmentally appropriate instrument with both parent- and youth-report forms available (Ravens-Sieberer & Bullinger, 1998). To help decide which therapeutic modality might be most helpful in reducing the depressive symptoms, the clinician considers Tandi's verbal ability educational level of the family, and cultural background, all of which suggest a good fit with cognitive behavioral or psychoeducational approaches. The clinician also decides to gather more information about family functioning to gauge the extent to which family dynamics and communication might helpful to address, perhaps indicating a greater emphasis on family-focused therapy.

Step 12. Solicit and Integrate Patient Preferences. As noted in the article, it makes sense to do "Step 12" whenever in the assessment sequence it would be helpful in making decisions about assessment or treatment. The clinician presents the initial formulation to the family, discussing how changes in Tandi's mood can offer a parsimonious explanation for the clinical picture emerging from the testing. During the discussion, the clinician is able to directly address the mother's concern about possible bipolar disorder, stating that the probability of bipolar disorder is currently quite low, and pointing to specific findings establishing the basis for that judgment. The clinician and family discuss several different options for treatment, ranging from "wait and see," through individual therapy for Tandi (involving supportive discussion combined with problem-solving and coping skills coaching), or family therapy, and antidepressant medication. Because no one in the immediate family has taken an antidepressant before, the clinician talks through the risks and benefits, providing the number needed to treat and the number needed to harm estimates for each approach. The family decides to try an approach combining some family psychoeducation with individual therapy for Tandi, holding the medication in abeyance because her depressive symptoms are still only mild to moderate, and thus the potential benefit seems lower compared to the potential for side effects and the family's hesitation about using medication.

Step 9. Measure Processes ("Dashboards, Quizzes and Homework"). Tandi and her mother download a mood charting app onto the mother's smartphone, and they use this to track both of their moods on a daily basis. This feeds directly into the mood monitoring and problem-solving skills that the clinician works to teach Tandi in individual sessions. The clinician also uses a sticker chart with Tandi to track the number of times each week that she tries new problem solving skills.

Step 10. Chart Progress and Outcome. In addition to regularly reviewing the mood charting and "homework" sticker chart, the clinician has Tandi and her mom repeat the CDI and CBCL after six sessions to see if there is measurable improvement on the primary outcomes. The family completes these a third time, along with repeating the quality of life measures, as they approach the termination session. The updated scores are compared to the "clinical significance" benchmarks as well as the baseline scores. Discussing the benchmarks helps the mother to reduce her sense of perfectionism, and allays her concerns that Tandi's moodiness might be a sign of bipolar disorder, by giving her a better appreciation for the behaviors that fall within typical functioning for Tandi's age.

Step 11. Monitor Maintenance and Relapse. During the termination session, the clinician and family review progress, celebrate their success, and plan for the future. This includes a discussion about the possibility of relapse. The clinician decides that this is important to discuss given the high rate of relapse for depression, and the fact that both early onset of depression and family history of mood disorder are risk factors that

increase Tandi's chances of remission. The clinician frames the potential for relapse as a possibility but emphasizes that Tandi and the family have mastered the skills to beat mood issues. The group discusses what would be warning signs of depression starting to recur, and they also make a list of situations that might increase stress and risk for relapse (such as getting a bad grade, losing a friend, getting very sick, or if the family were to relocate . . .). The list is framed as a set of "reminders" to check in on everyone's mood and coping when dealing with stressful situations. The clinician and mother also discuss warning signs that might raise concern about bipolar disorder, as both the family history and early onset suggest that if Tandi develops future mood issues, they are more likely to follow a bipolar spectrum course over the long term, even though she did not show signs of bipolar illness during this initial episode.

CHAPTER 15

ETHICS IN PSYCHOLOGICAL TESTING AND ASSESSMENT

Frederick T. L. Leong, Yong Sue Park, and Mark M. Leach

Since their early origins in the use of intelligence tests for placement of schoolchildren through the recent attention to high-stakes educational testing, psychological testing and assessment have remained controversial and complex topics. This controversy underscores the importance of addressing the ethical challenges in the use and application of tests and assessment in psychology. In this chapter, we begin with an overview of the various professional ethical standards that guide our work in this area. This section is followed by a more detailed review and discussion of the relevant sections of the American Psychological Association (APA) *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010). In this review, we also provide some guidance on the application of these ethical principles to the testing and assessment enterprise. Given the increasing cultural diversity of the U.S. population and the rise of globalization, we end with a discussion of some unique challenges in conducting testing and assessment cross-culturally.

There are also legal issues associated with testing and assessment in psychology, but these issues are not covered in this chapter because they are addressed elsewhere in this handbook (see Chapter 28, this volume, and Volume 2, Chapters 6 and 34). It is interesting to note that the U.S. Office for Human Research Protections highlights the differences between ethical principles and regulatory guidelines. *Ethical principles* refers to ethical values and principles aimed at the protection of human participants in research, whereas *regulatory guidelines* refers to a list of procedural dos and don'ts ("Distinguishing Statements of Ethical Principles and Regulatory Guidelines," 2011). The purpose of this chapter is to discuss the ethical values and principles in professional psychology as they pertain to testing and assessment.

PROFESSIONAL ETHICS

Ethics is a broad term that encompasses the commonly endorsed values of professional psychology (Groth-Marnat, 2006) and is the basis for ethics codes—rules and guidelines on appropriate behaviors for the purpose of protecting the public and the profession (Meara, Schmidt, & Day, 1996). In the United States, three major sources of ethics codes related to psychological testing and assessments are available: (a) the *Standards for Education and Psychological Testing* (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999), (b) the *Guidelines for Computer-Based Tests and Interpretations* (APA Committee on Professional Standards & Committee on Psychological Tests and Assessment, 1986), and (c) the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2010).

Standards for Education and Psychological Testing

In 1985, AERA, APA, and NCME collaborated to develop the *Standards for Education and Psychological Testing*—a set of standards pertaining to professional and technical issues of test development and use in education, psychology, and employment. The

DOI: 10.1037/14047-015
APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology, K. F. Geisinger (Editor-in-Chief)

Standards is organized in three sections: (a) Test Construction, Evaluation, and Documentation; (b) Fairness in Testing; and (c) Testing Applications. The *Standards* document was significantly revised in 1999 to contain a greater number of standards and updated to reflect changes in law and measurement trends, increased attention to diversity issues, and information on new tests and new uses of existing tests (AERA et al., 1999). An in-depth review of the *Standards* can be found in Chapter 13 of this volume.

Guidelines for Computer-Based Tests and Interpretations

With the increased use of, and concern for the lack of regulation of, psychological computer-based testing (CBT), APA's Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986) published the *Guidelines for Computer-Based Tests and Interpretations*, a set of 31 guidelines aimed at both test developers, to ensure the development of quality CBT products, and end users of these products, to ensure proper administration and interpretation of computer-based psychological tests (Schoenfeldt, 1989). More recently, the International Test Commission gave increased attention to CBT in its own set of CBT guidelines, adopted in 2005, titled the *International Guidelines on Computer-Based and Internet-Delivered Testing*. Similar to the objectives of the *Guidelines for Computer-Based Tests and Interpretations*, the general aim of the International Test Commission guidelines is to recommend standards for good practices for development and use of CBTs. The International Test Commission guidelines are organized along the following recommendations: (a) Give due regard to the technological issues in computer-based and Internet testing, (b) attend to quality issues in CBT and Internet testing, (c) provide appropriate levels of control over CBT and Internet testing, and (d) make appropriate provision for security and safeguarding privacy in CBT and Internet testing.

American Psychological Association Ethical Principles of Psychologists and Code of Conduct

APA adopted its first official code of ethics in 1952 in response to the field's increased professionalism and visibility after World War II (Fisher, 2009). Since then, the APA Ethics Code has been revised 10 times, with an amended version being adopted in 2010 by the APA Council of Representatives. The APA Ethics Code contains four major sections. The first section, Introduction and Applicability, delineates the rationale, scope and limitations, and applicability of the Ethics Code and describes the possible consequences and sanctions imposed on APA members and student affiliates who are found to have violated the standards of the Ethics Code. The second section, the Preamble, contains a statement of APA's purpose as a profession and delineates the various roles and responsibilities held by psychologists. The third section, General Principles, contains the five aspirational general principles of APA meant "to guide and inspire psychologists toward the very highest ethical ideals of the profession" (APA, 2010, p. 3): Beneficence and Nonmaleficence, Fidelity and Responsibility, Integrity, Justice, and Respect for People's Rights and Dignity. Finally, the fourth section, Ethical Standards, contains a set of 10 enforceable ethical standards by which psychologists are obligated to abide. Sanctions may be imposed on psychologists who violate these ethical standards. The ninth section of the Ethical Standards provides guidelines pertaining to the use of psychological tests and assessments (APA, 2010). In the next section, we discuss the APA ethical standards on assessments in greater detail as they apply to a variety of purposes and contexts in which psychological testing is conducted.

AMERICAN PSYCHOLOGICAL ASSOCIATION ETHICAL STANDARDS ON ASSESSMENTS

In the sections that follow, we highlight the 11 assessment standards associated with the APA Ethics Code. These standards have been found in the ethics codes of other countries, although the degree to which there is consistency differs based on a country's use of testing. In addition, other countries did include an additional standard not found in the APA Ethics Code (Leach & Oakland, 2007). The consistency found indicates that these standards

have international appeal and form the ethical foundation of test use and development.

Bases of Assessments

APA Ethical Standard 9.01, Bases for Assessments, stipulates that all oral and written opinions and conclusions made by psychologists be based on information and techniques grounded in the scientific and professional knowledge bases of professional psychology (Fisher, 2009). Adherence to the scientific and professional standards of the field builds public trust in the profession consistent with Principle B, Fidelity and Responsibility, of the APA Ethics Code. When psychologists' opinions and conclusions are not grounded in the scientific and professional standards, the probability that their opinions may mislead and potentially harm the clients and patients whom they serve is greater. Professional discernment applies to all phases of the testing and assessment process, even in the preassessment phase of planning and information gathering (Jacob & Hartshorne, 2006).

Scientific and professional bases. According to APA Ethical Standard 9.01a, psychologists are obligated to base their recommendations, reports, and diagnostic or evaluative statements on techniques supported by the scientific and professional standards of the field. Moreover, Ethical Standard 9.01b stipulates that opinions on individuals' psychological characteristics be drawn after an adequate examination is conducted on the basis of assessment procedures and tools that are consistent with the objective of the testing (e.g., that address the referral question), are sensitive to the cultural and linguistic characteristics of the examinee, are congruent with the examinee's level of competency to be administered the assessment, and have been shown to be valid and reliable. Psychologists are responsible for personally ensuring that the reliability and validity of the assessment tools and techniques they use are adequate. Furthermore, psychologists should base their conclusions and recommendations on assessments that have been demonstrated to be reliable and valid. Reliability and validity issues are discussed in greater depth in Chapters 2 and 4 of this volume.

Limitations of assessment results. When limitations to the reliability and validity of the assessment procedures and tools are found, psychologists should appropriately limit the nature and extent of their conclusions and recommendations and refrain from drawing conclusions that are not adequately supported. Another scenario to limit conclusions may arise when psychologists are unable to personally evaluate an individual for various reasons, such as an examinee's refusal to continue with assessment or an examinee's relocation during the course of assessment. In these situations, psychologists should make reasonable efforts, when appropriate and practical, to reach examinees for assessment and thoroughly document the outcome of these efforts (Ethical Standard 9.01b). When a personal evaluation is not practical, psychologists are obligated to limit the scope of their decisions and recommendations, in addition to delineating how the limited information influences the reliability and validity of their findings.

Cases may exist in which personal evaluation of an examinee is not warranted, such as when reviewing preexisting records in academic, legal, organizational, and administrative contexts or when examining secondary records provided by a third-party assessor, such as trainees or professionals with whom psychologists supervise or consult, respectively (Fisher, 2009; Knapp & VandeCreek, 2003). In these cases, psychologists should clearly explain that their conclusions and recommendations are based on a secondary analysis of information derived from alternate sources (Ethical Standard 9.01c).

Use of Assessments

Psychological testing applies to a wide range of purposes and contexts, which include but are not limited to screening applicants for job placement, diagnosing psychological disorders for mental health treatment, verifying health insurance coverage, conducting focus groups for market research, informing legal decisions and governmental policies, and developing measures to reliably measure personality characteristics (Aiken & Groth-Marnat, 2006; Fisher, 2009). According to the *Eighteenth Mental Measurements Yearbook* (Spies, Carlson, &

Geisinger, 2010), there are no less than 19 major categories of psychological tests and assessments.

APA Ethical Standard 9.02 pertains to the proper selection and use of psychological tests and assessments. The first component of this ethical standard stipulates that psychologists administer, adapt, score, interpret, and use psychological testing in the manner and purpose for which the selected tests and assessments were designed to be used as indicated by research (Ethical Standard 9.02a). Furthermore, psychologists should select and use tests or assessments with members of populations for whom adequate reliability and validity of the test scores has been established. If the reliability and validity of the test scores has not been examined or verified for a particular population, psychologists are obligated to describe the strengths and limitations of the interpretations and recommendations derived from the test or assessment results (Ethical Standard 9.02b). The third aspect of this ethical standard obligates psychologists to select tests and assessments that are appropriate to the language preference and competence of the individuals being assessed (Ethical Standard 9.02c).

Test selection and usage. Psychologists are responsible for selecting appropriate assessments for the intended purpose of the testing (Ethical Standard 9.02a). To guide the selection of appropriate tests and assessments, psychologists should have adequate knowledge of the theoretical bases and empirical evidence that support the validity and reliability of the tests or assessments; standardized administration and scoring procedures; approaches to interpreting the results; and the populations for which the assessment was normed and designed (Fisher, 2009; see Ethical Standard 9.07, Assessment by Unqualified Persons). Psychologists should also keep themselves updated on the most recent versions of the tests and assessments that they commonly use because testing and assessment procedures and parameters may change in light of theoretical advances and new research (see Ethical Standard 9.08, Obsolete Tests and Outdated Test Results). Finally, psychologists should select tests and assessments that have been empirically validated to be used in the specific contexts and settings in which the testing occurs.

Testing across diverse populations. According to Principles D (Justice) and E (Respect for People's Rights and Dignity) of the APA Ethics Code, psychologists strive to establish fair and equal access to and benefit of psychological contributions for all individuals and populations, which include but are not limited to diversity in age, gender, gender identity, race, ethnicity, culture, national origin, religion, disability, language, and socioeconomic status. Although psychological testing represents a unique contribution of professional psychology to benefiting larger society, ensuring the fair and equal access to and benefit of psychological testing has historically been challenging for the field. According to Reynolds (1982), the reliability and validity of test and assessment scores have predominately been established with White, middle-class samples and may not generalize well to other populations, especially those that represent a minority in the United States. This historical precedence conflicts with Ethical Standard 9.02b, which stipulates the selection and use of assessments that have been found to be adequately valid and reliable for drawing particular inferences for specific populations being assessed. When tests are administered across diverse populations, psychologists are obligated to select and use tests and assessments that have measurement equivalence in that the psychometric properties (i.e., measurement and structural models) have been shown to be equivalent or invariant between members of culturally different populations and those from the reference population for which the test and assessment scores were validated, normed, and found to be reliable (Schmitt, Golubovich, & Leong, 2010).

Testing and language. APA Ethical Standard 9.02c stipulates that psychologists select tests that are appropriate to be used with the language preferences and levels of competence of the individuals or groups being assessed. Thus, before selecting assessments, it is helpful for psychologists to gather information on examinees' cultural background (e.g., acculturation) and native and English language ability with regard to written, reading, and spoken language proficiencies (Jacob & Hartshorne, 2006; Takushi & Uomoto, 2001). According to Groth-Marnat (2009), literal translation of testing

and assessment materials and tools using the commonly implemented method of translation–back-translation may not be adequate because of cross-cultural differences in the conceptual interpretation of items, noncomparable idioms, and within-group differences in dialect and word usage. Furthermore, from an item response theory framework, literal translation of testing and assessment items from one language to another may change the properties of the items' difficulty, which may in turn diminish the measurement equivalence of tests or assessments. For these reasons, the psychometric properties of the original-language version of tests or assessments cannot be assumed to generalize to the alternate-language versions that were developed from a translation–back-translation method. More information on testing and language can be found in Volume 3, Chapter 26, of this handbook.

With regard to testing conducted in person (e.g., interviews) with linguistically different clients, psychologists may consider enlisting the services of a translator for interpretation purposes or consider referring clients to colleagues who have professional proficiency in the clients' language. Professional organizations may be useful resources for identifying and referring clients to professional colleagues with the appropriate linguistic background; for example, the National Association of School Psychologists maintains a directory of bilingual school psychologists that can be found on its website (http://www.nasponline.org/about_nasp/bilingualdirectory.aspx).

Informed Consent in Assessments

Before administering an assessment, psychologists are obligated to obtain from examinees, or their parents, guardians, or legal representatives, informed consent that includes an explanation of the nature and purpose of the assessment, fees, involvement of third parties (e.g., referral source), and limits of confidentiality (see Ethical Standard 3.10, Informed Consent). The informed consent stage of testing may also be the opportune time to provide examinees with an explanation of their rights as test takers. The Joint Committee on Testing Practices (1998) developed the *Rights and Responsibilities of Test Takers: Guidelines and Expectations* to inform test takers about and clarify expectations for the testing process. Because *consent* refers to examinees' legal status to autonomously decide whether to be assessed, informed consent must be communicated in a clear and comprehensible manner that is appropriate to the age of examinees and their mental abilities (Fisher, 2009).

As stipulated by Ethical Standard 9.03a, informed consent can be dispensed with in the following situations: when "(1) testing is mandated by law or governmental regulations; (2) informed consent is implied because testing is conducted as a routine educational, institutional or organizational activity; or (3) one purpose of the testing is to evaluate decisional capacity" (APA, 2010, p. 12). Even though informed consent is not required in these cases, psychologists are recommended to, when appropriate, continue to provide examinees with an explanation of the nature and purpose of the testing.

When assessing individuals younger than age 18 (i.e., minors), informed consent from parents or legal guardians is required because minors are viewed, from a legal standpoint, as being unable to make autonomous and well-informed decisions pertaining to psychological services. Thus, minors do not have the legal right to assent, consent, or object to a proposed psychoeducational assessment; however, it is recommended that minors be fully informed about the nature and purpose of the testing and assessment in a clear and understandable manner (Jacob & Hartshorne, 2006).

Nature and purpose of assessment. Informed consent in the assessment context includes an explanation of the nature and purpose of the test or assessment. Thus, psychologists are obligated to clearly explain how results will be used, the administration procedure, and possible benefits and risks or consequences of being assessed. With regard to informing examinees about the administration procedure, psychologists are advised to provide a general description of the procedure because foreknowledge of the testing may influence examinees' responses and thus alter the validity of the test or assessment results. Psychologists should also be sensitive to the possible risks and consequences of the testing, especially with regard to the negative

feelings that may be generated by the testing process. Some assessment topics or questions may elicit uncomfortable feelings in examinees, such as those that involve private or taboo topics (Groth-Marnat, 2009). Thus, psychologists, in most cases, should not pressure or force examinees to answer all questions, especially those that create undue discomfort or emotionally painful feelings.

Confidentiality and release of information. A core component of informed consent is explaining the limits of confidentiality. *Confidentiality* refers to a professional standard that requires psychologists to maintain the privacy of any assessment information unless disclosure is permitted or requested by examinees through a release of information. According to Ethical Standard 4.05, Disclosures, psychologists may breach confidentiality without examinees' permission when disclosure is mandated by law or when permitted by law for a valid purpose, such as to

> (1) provide needed professional services; (2) obtain appropriate professional consultations; (3) protect the client/patient, psychologist, or others from harm [e.g., danger to self and others, elder and child abuse]; and (4) obtain payment for services from a client/patient, in which instance disclosure is limited to only information that is necessary to obtaining the payment. (APA, 2010, p. 7)

In situations in which breach of confidentiality is necessary or legally mandated, psychologists should share only information that is necessary to accomplish the purpose of the disclosure in an effort to respect examinees' right to privacy.

Health Insurance Portability and Accountability Act and Family Educational Rights and Privacy Act. Because of the increased reliance on electronic databases to store client–patient information, psychologists are responsible for effectively protecting the confidentiality and security of the information contained in these databases (Aiken & Groth-Marnat, 2006). The Health Insurance Portability and Accountability Act (HIPAA) was established in 1996 to regulate the protection of protected health information. *Protected health information* refers to any information that

> (a) is created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse; and (b) relates to the past, present, or future physical or mental health or condition of any individual, the provision of health care to an individual, or the past, present, or future payment for the provision of health care to an individual. (Title 42, U.S.C. § 1320d)

Any health care provider who electronically transmits health information is considered a covered entity by HIPAA and must comply with HIPAA regulations. Within the informed consent, covered entities should provide examinees with a written document titled *Notice of Practice Practices;* this document contains a description of the examinee's rights, the legal duty to protect protected health information, and the routine uses and disclosures of protected health information. The Family Educational Rights and Privacy Act of 1974 pertains to issues of confidentiality and release of information in the educational setting. The act stipulates that assessment information and school records of students maintained by educational institutions that receive federal funding may be disclosed to others only with the written consent of the student examinees or their parents or legal guardians.

Language and use of interpretation services. Ethical Standards 9.03b and 9.03c refer to the psychologists' responsibility to provide informed consent in the language of the examinee or at a language proficiency level the examinee can reasonably understand. Psychologists may enlist the services of an interpreter when working with examinees who have limited English proficiency. When using interpreters, psychologists are responsible for ensuring that interpreters are not only competent in communicating the informed consent in a reasonable and understandable manner but also comply with the ethical standard on maintaining the confidentiality of examinees' identity, assessment results, and

test security (Fisher, 2009; Knapp & VandeCreek, 2003).

Release of Test Data

According to Fisher (2009), a growing trend in the legal system is toward affirming the autonomy of patients' access to their health care records, a trend that is consistent with Principle E, Respect for People's Rights and Dignity, of the APA Ethics Code, emphasizing self-determination. HIPAA stipulates that patients have the right to access, inspect, and receive copies of their medical and billing records on their request for the release of this information. Related to the assessment context, examinees or others identified in the release have the right, in most cases, to have access to their test data (Ethical Standard 9.04a). *Test data* refers to raw and scaled scores on the assessment items, any responses to test questions or stimuli, and psychologists' written notes or recordings of the testing.

Test data versus test materials. It is important to note the difference between test data and test materials. *Test materials* refers to test manuals, administration and scoring protocols, and test items. According to Ethical Standard 9.11, test materials do not need to be released pursuant to a client or patient request for test data because test materials are protected by copyright laws, and inappropriate release of such test materials is legally considered a breach of trade secrets (Groth-Marnat, 2009; Knapp & VandeCreek, 2003). However, when examinees' identifying information or responses are written on test materials, the test material is considered test data and may need to be released on examinees' request (Ethical Standard 9.04a). Thus, examiners are recommended, whenever possible, to record any identifying information and responses on a separate document from the actual test materials.

Potential misuse of test data. When examinees provide a release to request test data for themselves or identified others, it is important that psychologists explain the potential for test data to be misused if the people interpreting the test data do not have the proper qualifications to do so (see Ethical Standard 9.07, Assessment by Unqualified Persons). According to Ethical Standard 9.04a, psychologists may refrain from releasing test data to the examinees or others if the release may result in substantial harm resulting from misuse or misinterpretation of the test data. In these cases, psychologists are obligated to document the specific rationale for why they believe that the test data would result in substantial harm (Fisher, 2009).

Court order for test data. According to Ethical Standard 9.04b, psychologists are obligated to release test data when the disclosure is required by the law or court order. When release of test data is court mandated, Fisher (2009) recommended that psychologists seek legal counsel to determine the legitimacy of the request and ascertain their legal responsibility to release the test data. Another recommendation is that psychologists request the court for a protective order to prevent the inappropriate disclosure of the confidential test data and recommend that test data be reviewed by another health care professional who is qualified to provide appropriate and competent interpretations. Furthermore, psychologists are recommended to make reasonable efforts to notify examinees when test data are released to the court and to document these efforts (Fisher, 2009).

Test Construction

Ethical Standard 9.05, Test Construction, refers to test developers' responsibility to ensure that the development of tests and assessments incorporates appropriate psychometric procedures that are guided by the current scientific and professional knowledge of test design, standardization, validation, reduction or elimination of bias, and recommendations for use.

Standardization. Test developers are responsible for providing specific and clear guidelines to qualified test users with regard to the proper and standardized procedure for administering and scoring tests and assessments. Furthermore, test developers are responsible for specifying the scoring cutoffs and norms for the populations for which the tests and assessments were developed and intended to be used. Scoring norms are commonly found in norm-referenced tests, which allows for comparison of individual scores to the distribution of scores from

the reference group. It is important that the characteristics of the reference group sample are clearly described in the test or assessment manual and are representative of the population to which the test is targeted.

Validity. According to the *Standards for Educational and Psychological Tests* (AERA et al., 1999), *validity* is defined as the degree to which the theoretical basis for the assessment and accumulated empirical evidence support the intended interpretation of the scores for which the assessment was designed. In general, validity refers to the degree to which an assessment measures what it purports to measure. Several types of evidence are used to justify claims of validity, such as content-related evidence and criterion-related evidence. For an in-depth review, readers are referred to Chapter 4 in this volume.

Reliability. The *Standards for Educational and Psychological Tests* (AERA et al., 1999) stipulate that test developers are obligated to provide reliability estimates—the degree to which the assessment results are consistent over repeated administrations—of their tests and assessments. Jacob and Hartshorne (2006) recommended that reliability estimates be provided for each demographic subpopulation of the population for which the assessment was intended, such as for age groups and class levels. Several methods can establish the reliability of an assessment: internal consistency, test–retest, split-half test, and alternative-form comparisons. For an in-depth review, readers are referred to Chapter 2 in this volume.

Interpreting Assessment Results

Interpretations of test and assessment results influence the decisions and recommendations that are made in reference to the purpose of the testing (see Ethical Standard 9.02, Use of Assessments), such as diagnosing and informing treatment plans in clinical settings and educational placements in academic settings and determining employment selections and promotions. Interpretations should be based on proper administration of tests and assessments as outlined by the testing manual to ensure the interpretations are in line with the evidence to support the validity and reliability of the test or assessment scores (Fisher, 2009). It is the psychologist's responsibility to ensure that his or her interpretations of test or assessment results are useful and relevant to the purpose of the assessment and take into account various test factors, test-taking abilities, and other characteristics of individuals being assessed (Ethical Standard 9.06).

Interpretation of multiple sources. Interpretations of test and assessment results should not be derived from a simple, mechanical process that is based solely on the test or assessment scores, score cutoffs, or reliance on automated interpretations (Fisher, 2009; Groth-Marnat, 2009) but that takes into consideration a host of factors, including but not limited to examinees' characteristics, test-taking abilities, styles, issues of fatigue, perceptual and motor impairments, illnesses, language proficiencies, and cultural orientations (Fisher, 2009). Furthermore, Groth-Marnat (2009) recommended that psychologists base their interpretations on multiple sources of data, including behavioral observations, examinee background information, and other assessments. Often, testing is administered using an integrated battery of assessments, and inconsistent findings across the various assessments may result. In these situations, it is the psychologist's responsibility to analyze the contradictions and use his or her clinical and professional judgment to offer the most accurate and relevant interpretation in relation to the purpose of testing (Groth-Marnat, 2009).

Automated interpretations. There are many well-established, standardized assessments, such as the Minnesota Multiphasic Personality Inventory—2, for which one can receive a computer-generated automated interpretative report. Although these automated interpretations are based on a body of past empirical evidence and theoretical models, it is important to highlight that interpretations are not sophisticated enough to take into account examinees' unique characteristics and test-taking contexts. Thus, psychologists should not base their interpretations solely on automated interpretations but rather use automated interpretations as supplemental resources for integrated interpretations that take into consideration a host of other factors that may influence the testing.

Limitations of interpretations. According to Ethical Standard 9.06, Interpreting Assessment Results, psychologists are obligated to indicate any significant limitations of their interpretations, especially when the interpretations are not supported by the established validity and reliability of the test or assessment scores in making particular inferences. When interpretation of test or assessment scores is made outside their established validity and reliability, Fisher (2009) recommended that such interpretations be posed as hypotheses, rather than conclusions, to elucidate the limitations of such findings. Another limitation that needs to be indicated is when testing procedures and materials, evidence for validity and reliability, and score cutoffs and norms have become obsolete in the face of new research or changes in the populations for which tests and assessments were designed (see Ethical Standard 9.08, Obsolete Tests and Outdated Test Results).

Assessment by Unqualified People

APA Ethical Standard 9.07, Assessment by Unqualified Persons, warns against the promotion of psychological assessment techniques being used by unqualified people. Psychologists are obligated to ensure that testing is carried out by qualified individuals within the scope of their competence as indicated by their education and training background and past experiences (Fisher, 2009). Furthermore, qualified psychologists have knowledge of the nature and purpose of the assessments, their psychometric properties, standardized procedure for administration and scoring, proper interpretation of results, and assessment limitations (Groth-Marnat, 2009). Unqualified users may also include psychologists who are working with populations or problem areas that are outside the scope of their competencies (see Ethical Standard 2.01, Boundaries of Competence), such as working with culturally and linguistically different clients whom they are not multiculturally competent to serve.

Assessment by unqualified people may result in misdiagnosis of the examinees' presenting concerns and potentially result in psychological harm (Jacob & Hartshorne, 2006). Aiken and Groth-Marnat (2006) suggested that the unqualified use of assessments has greater consequences when assessing individuals (e.g., intelligence and personality assessments) as opposed to groups because misuse of assessment results can have direct negative consequences on people's livelihoods, such as being prescribed a treatment plan for an incorrect diagnosis or being placed at the wrong educational level or in the wrong job placement. In relation to Principle A, Beneficence and Nonmaleficence, of the APA Ethics Code, psychologists should be aware of the boundaries or limitations of their competence to prevent unqualified use of assessments and make appropriate referrals or seek supervision or consultation from specialists in these situations (Aiken & Groth-Marnat, 2006). Furthermore, psychologists are recommended to obtain access to or create a directory of local assessment specialists for referral purposes (Jacob & Hartsthorne, 2006).

Qualifications. According to Turner, DeMers, Fox, and Reed (2001), qualified use of assessments often includes graduate course work and supervised training experiences pertaining to the use of specific assessments. In 2002, the Psychological Assessment Work Group convened at the Competencies Conference: Future Directions in Education and Credentialing in Professional Psychology and identified a set of eight core competencies in psychological testing:

1. A background in the basics of psychometric theory.
2. Knowledge of the scientific, theoretical, empirical, and contextual bases of psychological assessment.
3. Knowledge, skill, and techniques to assess the cognitive, affective, behavioral, and personality dimensions of human experience with reference to individuals and systems.
4. The ability to assess outcomes of treatment/intervention.
5. The ability to evaluate critically the multiple roles, contexts, and relationships within which clients and psychologists function, and the reciprocal impact of these roles, contexts, and relationships on assessment activity.
6. The ability to establish, maintain, and understand the collaborative professional relationship that provides a context for all psychological activity including psychological assessment.

7. An understanding of the relationship between assessment and intervention, assessment as an intervention, and intervention planning.
8. Technical assessment skills that include: (a) problem and/or goal identification and case conceptualization, (b) understanding and selection of appropriate assessment methods including both test and non-test data (e.g., suitable strategies, tools, measures, time lines, and targets), (c) effective application of the assessment procedures with clients and the various systems in which they function, (d) systematic data gathering, (e) integration of information, inference, and analysis, (f) communication of findings and development of recommendations to address problems and goals, (g) provision of feedback that is understandable, useful, and responsive to the client, regardless of whether the client is an individual, group, organization or referral source. (Krishnamurthy et al., 2004, pp. 732–733)

The Psychological Assessment Workgroup also delineated core competencies of training programs in providing quality educational and training experiences for psychological testing.

Ethical responsibility for qualified use applies not only to individual psychologists but also to test developers with regard to the distribution of their test materials. Standards for qualified use have been established by test developers to prohibit unqualified users' access to test materials. Thus, test developers should include information on the required qualifications for use in the test's promotional materials and require end users to meet the minimum requirements to purchase and use their tests and assessments. Aiken and Groth-Marnat (2006) provided a sample qualification form for test developers that includes questions for the potential end user with regard to the purpose for using the test, area of professional expertise, level of training, specific courses taken, and quality control over test use (e.g., test security, appropriate tailoring of interpretations).

Assessment by trainees. Although APA Ethical Standard 9.07 stipulates that psychologists should not promote unqualified use of assessments, an exception is made for training purposes as long as trainees have adequate supervision while the assessments are provided. More specifically, for trainees to be qualified in administering tests or assessments, they must have been or concurrently be enrolled in a graduate-level course, practicum externship, or pre- or postdoctoral training program that provides training in the specific assessment that is being administered. In addition to the formal training, trainees must receive adequate supervision from a qualified user of the test or assessment. In cases in which unqualified trainees have not received sufficient training and supervision to administer the assessment, they must clearly inform examinees that the test or assessment is being administered for training purposes only and adequately describe the limitations of their assessment interpretations, conclusions, and recommendations (Fisher, 2009). It is important to note that when supervising psychologists sign their trainees' assessment reports, they are ultimately held responsible for the contents of the report (Jacob & Hartshorne, 2006).

Obsolete Tests and Outdated Test Results

Psychologists are prohibited from basing their decisions and recommendations on test data that are outdated for the test's current use (Ethical Standard 9.08a) and from tests and assessments that are obsolete and not useful for the current use (Ethical Standard 9.08b). Use of outdated test data is prohibited because examinees may have changed since the time of the prior assessment owing to such factors as maturational and developmental effects, development of new presenting problems, and changes in the environment (Fisher, 2009). When outdated test results are used, psychologists are obligated to provide an explanation for why outdated test data are used and to clearly communicate the limitations of such outdated information.

Old test data are often kept stored in outdated files or databases even after examiners no longer work at the testing location. In this situation, psychologists are recommended to prevent the misuse of outdated test results by taking reasonable steps to remove or destroy obsolete data and files. In cases in which clients or patients request that outdated test

data be sent to a new clinician who is currently providing services to them, psychologists are recommended to include a cover page detailing the limitations of outdated test results.

APA Ethical Standard 9.08 also stipulates that psychologists should not base their decisions and recommendations on use of obsolete assessments. According to Fisher (2009), tests developers often revise their assessments to reflect significant advances and changes in the theoretical constructs underlying the psychological characteristics being assessed; changes in the assessment's test item validity owing to various cultural, educational, linguistic, or societal influences; and shifts in the demographics of the target population, which in turn affect the standardized norms and score cutoffs. Use of obsolete tests may be applicable when long-term comparisons of test performance are needed, but psychologists are obligated to adequately describe the differences between test versions and explain the limitations of their comparisons when obsolete tests are used. According to Fisher (2009), the expense associated with updating to new versions is not an adequate ethical justification for using obsolete tests and assessments.

Test Scoring and Interpretation Services

APA Ethical Standard 9.09 applies to psychologists who provide test scoring and interpretation services. Within their promotional and other administrative materials (e.g., manuals), these psychologists are obligated to accurately describe the nature and purpose of the assessments, the basis for the standardized norms, and validity and reliability information for their assessment results and interpretations and to specify the qualifications for using the services. When interpretations and recommendations from assessment results are made, psychologists are obligated to provide the theoretical rationale and psychometric evidence for justifying their conclusions and to adequately explain the limitations of their interpretations and recommendations.

Ethical responsibility for the appropriate use of test scoring and interpretation services also applies to psychologists who are consumers of these services. These psychologists are obligated to select services that adequately provide evidence for the validity and reliability of their procedures for administering, scoring, and interpreting test and assessment results. Furthermore, psychologists using these services are obligated to have the qualifications and competence to ensure that the scoring and interpretations made by these services are consistent with APA Ethical Standard 9.06, Interpreting Assessment Results. When these services are used, the HIPAA Notice of Privacy Practices obligates psychologists to inform and obtain authorization from their clients or patients to permit the release of test or assessment information to these services.

Explaining Assessment Results

According to Ethical Standard 9.10, Explaining Assessment Results, psychologists are obligated to provide competent feedback to examinees, or to parents or legal guardians of minors, explaining any interpretations, decisions, and recommendations in relation to the purpose of testing. Groth-Marnat (2009) recommended that the feedback begin with a clear explanation of the rationale for testing, followed by the nature and purpose of the assessment, general conclusions drawn from assessment results, limitations, and common misconceptions or misinterpretations of assessment results. When examinees are minors, psychologists are obligated to provide the feedback to both examinees and their parents or legal guardians.

Sensitivity in the communication of assessment results. The *Standards for Educational and Psychological Tests* (AERA et al., 1999) stipulates that simple, clear, everyday language should be used when providing feedback so that the feedback is readily understood by its recipients. Psychologists should tailor their level of communication to recipients' personal characteristics, such as their educational and linguistic backgrounds, level of knowledge of psychological testing, and possible emotional reactions to the assessment results (Groth-Marnat, 2009). With regard to the possible emotional reactions generated by feedback, it may be helpful for psychologists to make available options for follow-up counseling to facilitate services for examinees who may need support in processing the feedback information. When providing

feedback on mental health status, Aiken and Groth-Marnat (2006) recommended that the least stigmatizing label be used to describe the examinees' psychological conditions or diagnoses.

Written reports. In addition to the oral feedback session, psychologists commonly provide written reports to examinees, or their referral source, regarding the assessment results, interpretations, and recommendations. Written reports should be centered on referral questions and the purpose of the testing and adequately describe the characteristics of the examinees and how they relate to the assessments used and the test situations (Aiken & Groth-Marnat, 2006). According to Jacob and Hartshorne (2006), written reports should be comprehensible to both professionals and nonprofessionals and should be written in a succinct, clear, and comprehensible manner while avoiding overgeneralizations (Aiken & Groth-Marnat, 2006). Psychologists are responsible for signing off on assessment reports only after ensuring the accuracy of the contents contained in the reports.

Maintaining Test Security

According to Ethical Standard 9.11, Maintaining Test Security, psychologists are obligated to maintain the security of test materials, which are defined as manuals, instruments, protocols, and test questions or stimuli. As noted in Ethical Standard 9.04, although examinees have the right to request and access test data, they do not have the right to access test materials for reasons related to threats to validity and copyright protection. For these reasons, test materials should be stored in a secure location, and only authorized and qualified individuals should have access to them. Furthermore, test materials, even sample items, should not be reprinted in any form, such as in newspapers and magazines, without the written consent of the test developers.

Threat to validity. A primary reason for the ethical obligation to maintain test security is the threat to test validity that is posed when individuals have access to test materials before administration of the test. Having foreknowledge of the test questions and answers may alter the psychometric properties of the test, including its standardized score cutoffs and norms and validity (Fisher, 2009). Furthermore, access to test materials before administration may increase the likelihood of some individuals manipulating their responses for purposes of malingering or obtaining an unfair advantage on a given assessment relative to others (Knapp & VandeCreek, 2003).

Copyright law. Pursuant to copyright protection laws, it is illegal and an ethical violation to reproduce test materials without obtaining permission from test developers or publishers. Maintaining test security allows for the protection of trade secrets and honors the terms of agreement made with the test publisher on obtaining access to the test materials (Groth-Marnat, 2009). With regard to HIPAA, which stipulates that examinees have the right to access their protected health information (e.g., test data), psychologists should separate, when appropriate, test materials from test data to protect the copyrighted test materials from being disclosed when releases of information are requested by clients or patients.

CROSS-CULTURAL ISSUES

Testing and assessment become inherently more complex when considering cross-cultural issues. Our position is that to be ethically and multiculturally competent when conducting testing and assessments, the psychologist should consider the client's cultural context. Approximately one third of the U.S. population consists of ethnic minorities, and when one includes the potential influence of other diverse groups (e.g., language, disability, socioeconomic status) on testing, it becomes evident that to be competent in testing and assessment requires much more than basic knowledge of test use.

All of the principles described in APA's (2010) *Ethical Principles of Psychologists and Code of Conduct* apply to cross-cultural testing, yet two are briefly highlighted that seem particularly salient. These are Principle D (Justice) and Principle E (Respect for People's Rights and Dignity). First, Principle D refers not only to equal access and fairness but to psychologists' ensuring that their biases, boundaries of competence, and level of expertise do

not influence their work and lead to unjust practices. Second, Principle E refers to respecting differences among individuals and cultural groups and the belief in autonomous self-determination. Unfortunately, sound ethical testing practices have not always been the norm when considering the history of the testing movement in psychology. Although progress in ethical testing practices has been made over the years and the field has improved significantly in the development, measurement, and implementation of testing with regard to culture, further developments are needed.

Psychological testing has made great strides in the understanding of psychological constructs, and it continues to do so. It also has a well-referenced history of bias against those who are not White, middle class, and male. The acceptance of the belief in universality, that the mainstream American experience is applicable to everyone, has long been at odds with a multicultural framework. This framework states that testing and assessment cannot be uniformly applied to all groups (Leong, Qin, & Huang, 2008). Using a simple example, readers would probably agree that assessing women if a test was normed on men or adults if a test was normed on elementary school-aged children would not be ethically appropriate. Similarly, there may be concerns about the application of tests primarily normed on the dominant group when considering use with nondominant group members. Consistent with many psychologists today, Burlew (2003) cautioned against taking a universal philosophical approach in that theories may not be transferable across cultures, that researchers are limited from developing alternative theories, that protective measures unique to a particular cultural group are neglected, and that any deviation from the universal perspective leads to a pathological or deviational view of nondominant outgroups. Only during the past few decades has research attention been given to the inclusion of diverse individuals and groups as they relate to the richness in understanding human behavior.

Etic Versus Emic

Validity from a cross-cultural perspective begins with knowledge of differences between etic and emic approaches to testing. Simply defined, etic approaches assess constructs across cultures, whereas emic approaches examine a construct within a particular culture. Understanding these validity issues is crucial when developing or using tests because tests are generally developed within a particular cultural context. Both etic and emic approaches are discussed in greater detail next, and examples from history are included to highlight ethical issues that have emerged.

Etic

Psychological testing has been at the forefront of controversy since the early part of the 20th century because of differences found among ethnic groups on a variety of tests, most notably intelligence tests. Imposed etics surrounding psychological assessment probably began with Galton's (1883/2003) treatise, "Inquiries Into Human Faculty and Its Development." This document led to the "mental test," which then helped launch psychology's version of the eugenics movement (Schultz & Schultz, 2011). Other psychologists such as Cattell, Goddard, and Terman were influential in launching intelligence and ability testing into conventional psychology. These famous psychologists, along with other equally as recognizable names such as Yerkes, were influential in putting forth testing practices that were unfavorable toward ethnic minorities, those of lower socioeconomic status, and others. More recently, Herrnstein and Murray's (1994) controversial book *The Bell Curve* revived the debate over the relationship among (primarily ethnic) groups and intelligence. Their thesis that ethnic minorities do not score well on tests of intelligence and achievement because of genetic and biological limitations harkens back to earlier testing history in psychology (for a review of the issues surrounding *The Bell Curve* and a rebuttal, see Jacoby & Glauberman, 1995).

Culturally appropriate and ethical test development has recently gained significant attention in the professional literature (e.g., Dana, 2005; Groth-Marnat, 2009). In this vein, to work toward competent, ethical, and culturally valid testing practices, psychologists and others have begun discussing test equivalence (or invariance). *Equivalence* refers to the

degree to which the parameters of a test's measurement model are comparable across groups (Cheung, van de Vijver, & Leong, 2011). Measurement equivalence is a prerequisite before one can make reasonable and ethical interpretations of the results across cultural groups. Historically, equivalence in psychological testing was omitted or significantly flawed given that many psychological tests were either normed on or developed in a framework of the dominant culture. Quite simply, using a psychological test that has not included a broader multicultural framework may introduce bias and is ethically dubious. It may be unethical because, among a myriad reasons, the psychologist is not acting competently and the foundation on which the tests were developed is flawed. More specifically, the APA Ethics Code acknowledges that ethical test use requires that the test be appropriate for the individual or group under investigation. Determination of whether a psychological instrument is valid for use with a particular cultural group is based on multiple factors, such as an individual's level of acculturation, translation of the instrument, language abilities, whether the construct measured with the instrument is consistent across cultures, and norm availability, among others. These can be accomplished through the assessment of four types of equivalence: linguistic, conceptual, metric, and functional (Leong, Leung, & Cheung, 2010).

Linguistic Equivalence

Linguistic equivalence, or translation equivalence, is primarily concerned with the translation of a psychological instrument and its application in another culture (Groth-Marnat, 2009). Brislin (1970) was one of the first to discuss the back-translation method, which involves translating an instrument into another language and then back-translating it into the primary language. The two versions are compared, and differences are resolved. Linguistic equivalence merely permits comprehensibility and does not, however, postulate about the instrument's validity. It is still a common translation method, although more recent procedures regarding the area of linguistic equivalence are expounded on in Hambleton, Merenda, and Spielberger (2005) and Volume 3, Chapter 26, of this handbook.

Conceptual Equivalence

Unfortunately, linguistic equivalence may be sufficient with some tests, but conceptual equivalence is also needed to behave in the highest ethical manner. Conceptual equivalence determines the degree to which a concept is consistent cross-culturally. This concept is more difficult to attain because what may be considered a similar concept between cultures may actually be a close proximity to it or interpreted differently altogether, resulting in conceptual variability. To decrease this variability, Usunier (1998) suggested that the translation process include multiple sources and target languages. Briefly, multiple native speakers independently develop words consistent with a concept, and a cross-cultural research team identifies the most commonly cited terms and back-translates them. Etic and emic conceptual dimensions are then determined (see also Leong et al., 2010).

Metric Equivalence

Metric equivalence is concerned with whether the psychometric properties of an instrument are consistent across cultural groups (Groth-Marnat, 2009). This type of equivalence is delineated into two categories, measurement invariance and structural invariance. Measurement invariance is related to variables' relationships to latent constructs, whereas structural invariance involves the actual latent variables themselves. Another way of considering the two is that measurement invariance is concerned with consistent matrices and scalar equivalence, for example, whereas structural invariance is concerned with whether the structural models, for example, are consistent across cultural groups. The more metric variability introduced, the greater the likelihood is that using the test across cultures is invalid and unethical.

Functional Equivalence

Functional equivalence addresses the idea that patterns of relationships between various constructs and a target measure are equivalent. If one construct in one culture does not function in the same manner in another culture, then variability is increased. For example, cognitive distortions may be associated with depression in one culture but not in another.

To test for cognitive distortions in one culture because of its cultural consideration as a common feature of depression in another culture could be inaccurate. To derive meaning and make interpretations from test results based on functional invariance could be considered unethical behavior (for a brief overview of strategies to offset measurement inequivalence, see Leong et al., 2008, 2010).

At least five ethical standards should be considered when evaluating tests without equivalence. We first consider a translated test developed in the English language and administered, for example, to an individual whose native language is Spanish. As indicated earlier, Ethical Standards 9.01, 9.02, and 9.06 are directly related to test use, and these three standards are central to linguistic equivalence. Standard 9.01, Bases for Assessments, states, "Psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, on information and techniques sufficient to substantiate their findings" (APA, 2010, p. 12). Without linguistic equivalence, for example, a simple translation without the back-translation, the psychologist is acting unethically because whether the translation is accurate is not clear. Whether the results can be used to substantiate the findings cannot be known.

Additionally, Ethical Standard 9.02, Use of Assessments, states,

> (a) Psychologists administer, adapt, score, interpret, or use assessment techniques, interviews, tests, or instruments in a manner and for purposes that are appropriate in light of the research on or evidence of the usefulness and proper application of the techniques.
>
> (b) Psychologists use assessment instruments whose validity and reliability have been established for use with members of the population tested. When such validity or reliability has not been established, psychologists describe the strengths and limitations of test results and interpretation.
>
> (c) Psychologists use assessment methods that are appropriate to an individual's language preference and competence, unless the use of an alternative language is relevant to the assessment issues. (APA, 2010, p. 12)

Standard 9.06, Interpreting Assessment Results, states,

> When interpreting assessment results, including automated interpretations, psychologists take into account the purpose of the assessment as well as the various test factors, test-taking abilities, and other characteristics of the person being assessed, such as situational, personal, linguistic, and cultural differences, that might affect psychologists' judgments or reduce the accuracy of their interpretations. They indicate any significant limitations of their interpretations. (APA, 2010, p. 13)

Two general competence standards are applicable as well. Standard 2.01(b), Boundaries of Competence, states,

> Where scientific or professional knowledge in the discipline of psychology establishes that an understanding of factors associated with age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, or socioeconomic status is essential for effective implementation of their services or research, psychologists have or obtain the training, experience, consultation, or supervision necessary to ensure the competence of their services, or they make appropriate referrals. (APA, 2010, p. 5)

Finally, Ethical Standard 2.04, Bases for Scientific and Professional Judgments, indicates that psychologists should use only the best scientific and professional methods in their work. Unless linguistic equivalence is achieved to the highest standard possible, then the psychologist is in danger of failing to measure up to this standard.

Emic

The emic approach to test use has historically been at odds with the etic approach. An emic approach is consistent with an indigenous approach in that it is culture specific. In essence, tests are developed for particular groups under investigation without the need to expand them to other groups. It is limited in that a narrow understanding of a particular group does not increase one's broader understanding of psychological processes common to all individuals. However, we believe that more culture-specific tests are needed to gain a more robust understanding of diverse groups. Further theory development integrating both mainstream and indigenous psychologies will occur through increased development and recognition of culturally specific tests (Morris, Leung, Ames, & Lickel, 1999).

Although development and assessment of culture-specific tests has increased, a combined etic–emic approach to testing and assessment has recently received increased attention. Constructs derived indigenously are combined with local interpretations of universal constructs to offer a comprehensive measurement instrument relevant to a particular cultural context. Using an international example, the Chinese Personality Assessment Inventory (Cheung et al., 1996) is an instrument that combines both etic and emic perspectives. Local expressions of Chinese culture from a variety of China's regions served as the foundation for both culturally relevant and universal constructs. It overlaps with the Big Five scales, but a relational factor also emerged that is consistent with collectivistic cultures. It has great promise for future test development owing to the methodological approach taken, and it has been used in multiple regions of the world (Leong et al., 2010).

Additional Ethical Test Practices and Diversity

The APA Ethics Code has ethical practice standards that have relevance to diverse communities. These standards should be considered from a contextual framework. Some were mentioned earlier when discussing equivalence issues and two others are highlighted next. Although not explicitly stated, Standard 9.07, Assessment by Unqualified Persons, applies to those lacking sufficient cultural competence. For example, even culturally competent psychologists should be cognizant that not everyone with whom they work has the same level of cultural expertise. Colleagues should not be asked to administer, score, and interpret tests without proper understanding of their cultural context. When considering culture, this standard is also related to Standard 9.02, Use of Assessments. Standard 9.10, Explaining Assessment Results, becomes particularly salient when considering individuals whose second or third language is English and those who are unfamiliar with the purpose of testing. This standard is also related to Standard 9.03, Informed Consent in Assessments.

Although they are discussed in terms of school psychology assessments, Jacob and Hartshorne (2007) perhaps best summarized the ethical issues that arise from conducting broader culturally valid assessments. They determined that assessments should be multifaceted, comprehensive, fair, valid, and useful. As psychologists' understanding of cultural tests and assessments increases and becomes integrated into test development and use, they will feel comfortable using tests that cover these five issues, leading to greater ethical and cultural competence.

References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Upper Saddle River, NJ: Pearson Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (Rev. ed.). Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Association.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct (2002, amended June 1, 2010)*. Retrieved from www.apa.org/ethics/code/index.aspx

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments. (1986).

Guidelines for computer-based tests and interpretations. Washington, DC: American Psychological Association.

Brislin, R. (1970). Back translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185–216. doi:10.1177/135910457000100301

Burlew, A. K. (2003). Research with ethnic minorities: Conceptual, methodological, and analytical issues. In G. Bernal, J. E. Trimble, A. K. Burlew, & F. T. L. Leong (Eds.), *Handbook of racial and ethnic minority psychology* (pp. 179–197). Thousand Oaks, CA: Sage. doi:10.4135/9781412976008.n9

Cheung, F. M., Leung, K., Fan, R., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese Personality Assessment Inventory (CPAI). *Journal of Cross-Cultural Psychology, 27,* 181–199. doi:10.1177/0022022196272003

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66,* 593–603.

Dana, R. H. (2005). *Multicultural assessment: Principles, applications, and examples.* Mahwah, NJ: Erlbaum.

Distinguishing statements of ethical principles and regulatory guidelines. (2011). Retrieved from http://med.brown.edu/fogarty/codes.htm#disting

Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g.

Fisher, C. B. (2009). *Decoding the ethics code: A practical guide for psychologists* (2nd ed.). Thousand Oaks, CA: Sage.

Galton, F. (2003). Inquiries into human faculty and its development. In M. P. Munger (Ed.), *The history of psychology: Fundamental questions* (pp. 232–247). New York, NY: Oxford University Press. (Original work published 1883)

Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: Wiley.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, NJ: Erlbaum.

Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191.

International Test Commission. (2005). *International guidelines on computer-based and Internet-delivered testing.* Retrieved from http://www.intestcom.org/itc_projects.htm

Jacob, S., & Hartshorne, T. S. (2006). *Ethics and law for school psychologists* (5th ed.). Hoboken, NJ: Wiley.

Jacoby, R., & Glauberman, N. (1995). *The bell curve debate: History, documents, opinions.* New York, NY: Random House.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations.* Retrieved from http://www.apa.org/science/programs/testing/rights.aspx

Knapp, S., & VandeCreek, L. (2003). An overview of the major changes in the 2002 APA ethics code. *Professional Psychology: Research and Practice, 34,* 301–308. doi:10.1037/0735-7028.34.3.301

Krishnamurthy, R., VandeCreek, L., Kaslow, N. J., Tazeau, Y. N., Miville, M. L., Kerns, R., . . . Benton, S. A. (2004). Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology, 60,* 725–739. doi:10.1002/jclp.20010

Leong, F. T. L., Leung, K., & Cheung, F. M. (2010). Integrating cross-cultural psychology research methods into ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology, 16,* 590–597. doi:10.1037/a0020127

Leong, F. T. L., Qin, D., & Huang, J. L. (2008). Research methods related to understanding multicultural concepts. In J. K. Asamen, M. L. Ellis, & G. L. Berry (Eds.). *Handbook of child development, multiculturalism, and media* (pp. 63–80). Thousand Oaks, CA: Sage.

Meara, N. M., Schmidt, L. D., & Day, J. D. (1996). Principles and virtues: A foundation for ethical decisions, policies, and character. *Counseling Psychologist, 24,* 4–77. doi:10.1177/0011000096241002

Morris, M. W., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Integrating emic and etic insights about culture and justice judgment. *Academy of Management Review, 24,* 781–796.

Reynolds, C. R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 192–227). Baltimore, MD: Johns Hopkins University Press.

Schmitt, N., Golubovich, J., & Leong, F. T. L. (2010). Impact of measurement invariance on construct correlations, mean differences and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment.* Advance online publication. doi:10.1177/1073191110373223

Schoenfeldt, L. F. (1989). Guidelines for computer-based psychological tests and interpretations. *Computers in Human Behavior, 5,* 13–21.

Schultz, D. P., & Schultz, S. E. (2011). *A history of modern psychology* (10th ed.). Belmont, CA: Wadsworth.

Spies, R. S., Carlson, J. F., & Geisinger, K. F. (2010). *Eighteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Takushi, R., & Uomoto, J. M. (2001). The clinical interview from a multicultural perspective. In

L. A. Suzuki, J. G. Ponterotto, & P. J. Meller (Eds.), *Handbook of multicultural assessment* (2nd ed., pp. 47–66). San Francisco, CA: Jossey-Bass.

Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist, 56*, 1099–1113. doi:10.1037/0003-066X.56.12.1099

Usunier, J. C. (1998). *International and cross-cultural management research.* London, England: Sage.

Practice Parameter for the Assessment and Treatment of Children and Adolescents With Autism Spectrum Disorder

Fred Volkmar, MD, Matthew Siegel, MD, Marc Woodbury-Smith, MD, Bryan King, MD, James McCracken, MD, Matthew State, MD, PhD, and the American Academy of Child and Adolescent Psychiatry (AACAP) Committee on Quality Issues (CQI)

Autism spectrum disorder is characterized by patterns of delay and deviance in the development of social, communicative, and cognitive skills that arise in the first years of life. Although frequently associated with intellectual disability, this condition is distinctive in its course, impact, and treatment. Autism spectrum disorder has a wide range of syndrome expression and its management presents particular challenges for clinicians. Individuals with an autism spectrum disorder can present for clinical care at any point in development. The multiple developmental and behavioral problems associated with this condition necessitate multidisciplinary care, coordination of services, and advocacy for individuals and their families. Early, sustained intervention and the use of multiple treatment modalities are indicated. J. Am. Acad. Child Adolesc. Psychiatry, 2014;53(2):237–257. **Key Words:** autism, Practice Parameters, guidelines, developmental disorders, pervasive developmental disorders

Since the first Practice Parameter for the Assessment and Treatment of Children, Adolescents, and Adults with Autism and Other Pervasive Developmental Disorders[1] was published, several thousand research and clinical articles have appeared and the diagnostic criteria for autism have changed. This Parameter revision provides the opportunity to update the previous version and incorporate new research. Because the extant body of research was performed under the *DSM-IV-TR* diagnostic schema, the evidence will be presented using that terminology. This Parameter is applicable to evaluation of children and adolescents (≤17 years of age) but often will have some relevance to adults. This document presumes basic familiarity with aspects of normal child development and child psychiatric diagnosis and treatment. Unless otherwise noted, the term *child* refers to adolescents and younger children, and *parents* refers to the child's primary caretakers regardless of whether they are the biological or adoptive parents or legal guardians.

CME This article can be used to obtain continuing medical education (CME) at www.jaacap.org.

METHODOLOGY

The first version of this Parameter was published in 1999. For this revision, the literature search covered the period from 1991 to March 19, 2013 using the PubMed, PsycINFO, Cochrane, and CINAHL (EBSCO) databases. The initial searches were inclusive and sensitive. Search terms were a combination of MeSH headings and keywords, and the MeSH headings were adjusted to terms used by PsycINFO and CINAHL by using their thesauri.

In PubMed the MeSH terms *autistic disorder, childhood development disorders—pervasive, Asperger**, and *Rett** and the keyword *autism* were searched. The initial search yielded 20,807 results. Then, the results were limited to English, human, *all child (0 to 18 years)*, and 1991 to March 19, 2013. Additional limits included classic article, clinical trial, comparative study, controlled clinical trial, evaluation studies, guideline, historical article, meta-analysis, practice guideline, multicenter study, randomized controlled trial, review, twin study,

and validation studies. The refined PubMed search yielded 3,613 articles.

In the PsycINFO database subject headings (focused) of *autism, autistic thinking, pervasive developmental disorders, retts syndrome, aspergers,* and keyword *autism* were searched. The initial search returned 24,875 articles and was then limited to English, *childhood: birth to age 12yrs, adolescence: age 13-17 yrs, peer reviewed journal,* and 1991 to March 19, 2013. The refined PsycINFO search yielded 9,583 articles.

In the Cochrane Database of Systematic Reviews, keywords of *autism, autistic, rett*, asperger*,* or (*pervasive and disorder** and *develop**) were searched without additional limits. The Cochrane search yielded 95 articles. An additional 517 articles were retrieved from the CINAHL database, after excluding Medline articles, by searching *autistic disorder, autism, asperger syndrome, child development disorders, pervasive,* and *rett syndrome.*

A total of 13,808 articles were identified and exported to the EndNote reference management program. After removing duplicate references, the resulting yield from the comprehensive search was 9,581 articles.

The titles and abstracts of all articles were reviewed. Studies were selected for full text review based on their place in the hierarchy of evidence (e.g., randomized controlled trials), quality of individual studies, and generalizability to clinical practice. The search was augmented by review of articles nominated by expert reviewers and further search of article reference lists and relevant textbook chapters. A total of 186 articles were selected for full text examination.

CLINICAL PRESENTATION AND COURSE

Autism was first described in 1943 by Kanner[2] who reported on 11 children with an apparently congenital inability to relate to other people but who were quite sensitive to change in the nonsocial environment. Kanner emphasized that the lack of interest in people was in stark contrast to the profound social interest of normal infants. He also observed that when language developed at all, it was marked by echolalia, pronoun reversal, and concreteness. The children also exhibited unusual, repetitive, and apparently purposeless activities (stereotypies). Autism was initially believed to be a form of childhood psychosis, but, by the 1970s, various lines of evidence made it clear that autism was highly distinctive. By 1980, autism was officially recognized as a diagnosis in *DSM-III.*[3]

Under *DSM-IV-TR,* the diagnosis of autism required disturbances in each of 3 domains: social relatedness, communication/play, and restricted interests and activities with onset by 3 years of age.[4] The disturbance in social relatedness is striking and includes marked impairment in nonverbal communication, peer relationships, and social-emotional reciprocity. Impairments in communication include a delay or total lack of spoken language (without an attempt to compensate through other means) or, for verbal individuals, a marked difficulty in the ability to sustain or initiate conversation, stereotyped and repetitive (or idiosyncratic) language, and lack of developmentally appropriate make-believe or social play. Impairment in interests and activities includes encompassing preoccupations, adherence to apparently nonfunctional routines or rituals, stereotypies and motor mannerisms, and persistent preoccupation with parts of objects.

There is variability in the age at which children may present the features essential for this diagnosis.[5] Preschool children with autism typically present with marked lack of interest in others, failures in empathy, absent or severely delayed speech and communication, marked resistance to change, restricted interests, and stereotyped movements. Common parental concerns include a child's lack of language, inconsistencies in responsiveness, or concern that the child might be deaf. In children with autism, social and communication skills usually increase by school age; however, problems dealing with change and transitions and various self-stimulatory behaviors (sometimes including self-injury) also may become more prominent during this time.[6] In adolescence, a small number of individuals with autism make marked developmental gains; another subgroup will behaviorally deteriorate (e.g., tantrums, self-injury, or aggression). Children and adolescents with autism have an increased risk for accidental death (e.g., drowning).[7] Predictors of ultimate outcome include the presence of communicative speech by 5 years of age and overall cognitive ability (IQ). Evidence that earlier detection and provision of services improves long-term prognosis makes early diagnosis particularly important.[8]

The *DSM-IV-TR* category of pervasive developmental disorders included autistic disorder, Rett's disorder, Asperger's disorder, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS). Rett's disorder was described by Andreas

Rett in 1966 in a series of girls with unusual hand washing/wringing stereotyped mannerisms. In most cases, Rett's disorder is caused by mutations in the gene *MeCP2* (methyl-CpG-binding protein 2).[9] Head circumference and development are normal at birth and during infancy. Before 4 years of age, head growth decelerates, purposeful hand movements are lost, and characteristic stereotyped hand movements (wringing or washing) develop.[10] The central role of *MeCP2* mutations in this disorder makes it clear that boys may carry the same mutations that lead to the full syndrome in girls, but with differing clinical manifestations ranging from fatal encephalopathy[11] to progressive but nonfatal developmental disorder[12] to nonspecific X-linked intellectual disability.[13]

Childhood disintegrative disorder (CDD) was first described by Theodor Heller in 1908.[14] This condition is characterized by a period of at least 2 years of normal development, followed by a marked deterioration and clinically significant loss of at least 2 skills in the areas of receptive or expressive language, social skills, toileting skills, play, or motor skills.[14] The onset of CDD is highly distinctive, typically occurring at 3 to 4 years of age and can be gradual or abrupt. Sometimes parents report that the child experienced a period of anxiety or dysphoria before onset of CDD symptoms. Once established, CDD resembles autism in clinical features,[14] but the outcome is poor. The child typically becomes mute or, at best, regains limited speech.

Asperger's disorder was described in 1944 but not officially recognized until *DSM-IV*. Unlike children with autism, individuals with Asperger's disorder do not present with delays in language acquisition or with unusual behaviors and environmental responsiveness during the first years of life. Consequently, parents often have no concerns about their child's early development.[15] Asperger originally described children who were precocious in learning to talk but who then talked in a formal, pedantic, 1-sided way, often about a topic of circumscribed interest.[16] Social difficulties arise due to this idiosyncratic, 1-sided social style. The outcome in Asperger's disorder generally appears to be better than that for autism, although this may, in part, relate to better cognitive and/or verbal abilities.[8,15]

The term pervasive developmental disorder not otherwise specified (PPD NOS) (also sometimes termed atypical PDD or atypical autism) encompasses subthreshold cases on the autism spectrum, e.g., cases in which full criteria for one of the explicitly defined PDDs are not met, but the child has problems in social interaction and some difficulties in communication or restricted patterns of behavior. Although studies are limited, individuals with PDD-NOS typically have been characterized as less impaired, having fewer repetitive behaviors, and having a better prognosis than persons with autism.[16]

DSM-IV-TR to *DSM-5*

Because there was little evidence to support reliable and replicable diagnostic differences among the various *DSM-IV-TR* PDDs,[17] the *DSM-5* workgroup on neurodevelopmental disorders subsumed the prior categories under the new diagnosis of autism spectrum disorder (ASD) in the *DSM-5*. Diagnostic domains were reduced from 3 to 2, focusing on social communication and interaction deficits and restricted, repetitive patterns of behaviors and interests. The strict requirement for onset before 3 years of age was changed to onset in the early developmental period, the occurrence of potential sensory abnormalities was incorporated, and a severity scale for impairments in each of the 2 core domains was included. Diagnostic reporting now includes specifiers that may enhance descriptive subtyping of the population, including specifiers for the presence or absence of intellectual impairment, language impairment, catatonia, and known medical, genetic, or environmental factors. The new criteria allow for a history of symptoms that may not be present currently, recognizing that through intervention or normal development some children with autism no longer present some symptoms later in life. It will be some years before the implications of these changes for autism prevalence and other facets of assessment and treatment can be fully assessed.

EPIDEMIOLOGY

Many studies, mostly conducted outside the United States, have examined the prevalence of autism or, less commonly, ASD or PDDs.[17] Of the approximately 36 surveys of autism available, prevalence estimates for autistic disorder range from 0.7 in 10,000 to 72.6 in 10,000.[18] The variability in estimates reflects different factors, including changes in definition. When the 18 surveys conducted since the introduction of the *DSM-IV* criteria are considered, estimates ranging from 10 in 10,000 to 16 in 10,000, with a median prevalence of 13 in 10,000, are obtained.[18] The most recent study by the Centers for Disease

Control and Prevention estimated the prevalence of ASD in the United States as 11.3 in 1,000.[19] Contrary to popular perception, data from 7 surveys suggest that rates of Asperger's disorder are in fact *lower* than for typical autism (2.6 in 10,000 or one fifth as common as typical autism).[18]

Recent observations of higher rates of autism have led to concern that the prevalence of this disorder may be increasing. Various factors may contribute to an apparent increase,[20] such as differences in diagnostic criteria and diagnostic practices, the age of children screened, and the location of the study (see Fombonne[18] for discussion).

Autism is approximately 4 times more common in males than in females, but females with autism tend to have more severe intellectual disability. Although the original report by Kanner[2] suggested a predominance of autism in more educated families, subsequent work has not shown this. Current approaches to the diagnosis of ASD appear to work well internationally and cross-culturally,[3] although cultural aspects of the condition have not received much attention.[21] Within the United States, there may be underdiagnosis in some circumstances (e.g., in disadvantaged inner-city children).[22]

ETIOLOGY

Neurobiology

Electroencephalographic (EEG) abnormalities and seizure disorders are observed in as many as 20% to 25% of individuals with autism.[23] The high rates of epilepsy suggest a role for neurobiologic factors in autism.[13,24,25] The number of areas affected by autism suggests that a diverse and widely distributed set of neural systems must be affected. Although various theories have posited potential loci for difficulties, definitive data are lacking. Postmortem studies have shown various abnormalities, particularly within the limbic system.[25] Functional magnetic resonance imaging procedures have identified difficulties in tasks involving social and affective judgments and differences in the processing of facial and nonfacial stimuli.[26] Structural magnetic resonance imaging has shown an overall brain size increase in autism, and diffusion tensor imaging studies have suggested aberrations in white matter tract development.[27] One of the most frequently replicated neurochemical findings has been the elevation of peripheral levels of the neurotransmitter serotonin. The significance of this finding remains unclear. A role for dopamine is suggested given the problems with overactivity and stereotyped mannerisms and the positive response of such behaviors to neuroleptic medications.[28]

During the past decade, much concern has focused on vaccines as a possible postnatal environmental cause for ASD, with the concern focused on the possibility that the measles-mumps-rubella vaccine may cause autism or that thimerosal (a mercury-containing preservative now removed from all single-dose vaccines) might do so.[29] The preponderance of available data has not supported either hypothesis (see Rutter[30] for a review). However, a possible role of the immune system in some cases of autism has not been ruled out.[31]

Neuropsychological correlates of ASD include impairments in executive functioning (e.g., simultaneously engaging in multiple tasks),[32] weak central coherence (integrating information into meaningful wholes),[33] and deficits in theory-of-mind tasks (taking the perspective of another person).[34]

Familial Pattern and Genetic Factors

The high recurrence risk for autism in siblings and even higher concordance for autism in identical twins has provided strong support for the importance of genetic factors.[30] Higher rates of autism are consistently noted in siblings of affected children. Recurrence risk has typically been cited at 2% to 10%, but a recent prospective longitudinal study has reported a rate of 18.7% when the broad autism spectrum is considered.[35] Identified risk factors for ASD appear to include closer spacing of pregnancies, advanced maternal or paternal age, and extremely premature birth (<26 weeks' gestational age).[36-38] In addition, high rates of learning/language problems and social disability and a possible increase in the risk for mood and anxiety disorders has been noted in family members.

It is now clear that multiple genes are involved in autism.[30,39] Over the past several years, studies have supported a role for common (present in >5% of the general population) and rare genetic variations contributing to autism.[40] The rate of progress in gene discovery has been increasing rapidly over the past several years and these results are already beginning to influence clinical practice with regard to genetic testing, as noted below.[41]

DIFFERENTIAL DIAGNOSIS

ASD must be differentiated from specific developmental disorders (including language disorders),

sensory impairments (especially deafness), reactive attachment disorder, obsessive-compulsive disorder, intellectual disability, anxiety disorders including selective mutism, childhood-onset schizophrenia, and other organic conditions.

A diagnosis of autism is made when the requisite *DSM-5* symptoms are present and other disorders have been adequately ruled out. In autism it is typical for parents to report that there was no period of normal development or that there was a history of unusual behaviors (e.g., the child seemed too good and undemanding as an infant). Less commonly, a period of apparently normal development is reported before a regression (loss of skills). The topic of regression in autism remains an active area of current investigation. Developmental regression is typical in Rett syndrome but also can be observed in other conditions (e.g., childhood-onset schizophrenia or degenerative CNS disorders).

Developmental language disorders have an impact on socialization and may be mistaken for an ASD. The distinction is particularly difficult in preschool children. However, 2 behaviors have been reported to consistently differentiate autistic children from language-impaired peers at 20 and 42 months of age, namely pointing for interest and use of conventional gestures.[42] Similarly, differentiating mild to moderate developmental delay from ASD may be difficult, particularly when evaluating the younger child (see Chawarska and Volkmar[42] for a detailed discussion). One study identified some items on the Autism Diagnostic Interview that differentiated between these 2 groups at 24 months, especially directing attention (showing) and attention to voice (Table 1).[43-56] At 36 months, 4 items correctly classified all subjects: use of other's body, attention to voice, pointing, and finger mannerisms. From 38 to 61 months, children with autism were more likely to show impaired nonverbal behaviors (such as eye contact) to regulate social interaction. In childhood, there may be diagnostic overlap between ASD and attention-deficit/hyperactivity disorder, making the differential diagnosis difficult.[57,58]

Children with reactive attachment disorder may exhibit deficits in attachment and therefore inappropriate social responsivity, but these usually improve substantially if adequate caretaking is provided. Obsessive-compulsive disorder has a later onset than ASD, is not typically associated with social and communicative impairments, and is characterized by repetitive patterns of behavior that are ego dystonic. Symptoms that characterize anxiety disorders, such as excessive worry, the need for reassurance, the inability to relax, and feelings of self-consciousness, are also seen in ASD, particularly in higher functioning individuals. However, the 2 conditions can be differentiated by the prominent social and communicative impairments seen in ASD but not anxiety disorders, and the developed social insight of children with anxiety disorders, which is not seen in ASD. Differentiating childhood schizophrenia from autism can be difficult, because they are characterized by social impairments and odd patterns of thinking. However, florid delusions and hallucinations are rarely seen in autism.

COMORBIDITIES

Given difficulties in communication (e.g., mutism) and cognitive impairment, issues of comorbidity in ASD can be quite complex. The process of diagnostic overshadowing (the tendency to fail to diagnose other comorbid conditions when a more noticeable condition is present) may occur.[59] Attempts to determine comorbidity prevalence in ASD have been hampered by methodologic issues, although most studies have shown increased rates of anxiety and attentional disorders.[60]

In most epidemiologically based samples of persons with autistic disorder, approximately 50% exhibit severe or profound intellectual disability, 35% exhibit mild to moderate intellectual disability, and the remaining 20% have IQs in the normal range.[18] For children with autistic disorder, verbal skills are typically more impaired than nonverbal skills. For children with Asperger's disorder, the reverse pattern is sometimes observed and the profile of nonverbal learning disability may be present.[61] Clearly, intellectual impairment is not an essential diagnostic feature of autism, and thus it is necessary and important for the diagnosis of intellectual disability to be made.

A range of behavioral difficulties can be observed in ASD, including hyperactivity, obsessive-compulsive phenomena, self-injury, aggression, stereotypies, tics, and affective symptoms. The issue of whether these qualify as additional disorders is complex.[3] Affective symptoms are frequently observed and include lability, inappropriate affective responses, anxiety, and depression. Impairments in emotion regulation processes can lead to under- and

TABLE 1 Summary of Selected Assessment Instruments for Autism Spectrum Disorder[a]

Scale (see legend)	Uses	Age Range	Method of Administration	Population Studied	Scale characteristics	Reference
ABC	screening	children	parent rated	AD	57 items, scale 1-4	Krug *et al.*, 1980[43]
CARS	screening	children	clinician rated	AD	15 items, scale 1-4	Schopler *et al.*, 1980[44]
M-CHAT	screening	toddlers	parent rated	AD	23 items, yes/no	Robins *et al.*, 2001[45]
CSBS-DP-IT-Checklist	screening	toddlers	parent rated	AD	24 items	Wetherby *et al.*, 2008[46]
ASQ	screening	child/adult	parent rated	AD/AspD	40 items, yes/no	Berument *et al.*, 1999[47]
AQ	screening	child/adult	self or parent rated	AspD	50 items, scale 0-3	Baron-Cohen *et al.*, 2001[48]
CAST	screening	4-11 years	parent rated	AspD	37 items, yes/no	Scott *et al.*, 2002[49]
ASDS	screening	5-18 years	parent or teacher rated	AspD	50 items, yes/no	Myles *et al.*, 2000[50]
GADS	screening	3-22 years	parent or teacher rated	AspD	32 items, scale 0-3	Gilliam, 2001[51]
ASDI	screening	child/adult	interview + clinician rated	AspD	50 items, yes/no	Gillberg *et al.*, 2001[52]
SRS	screening	4-18 years	parent or teacher rated	AspD	65 items, scale 1-4	Constantino *et al.*, 2003[53]
ADI	diagnostic	child/adult	interview + clinician rated	AD/AspD	see text	Lord *et al.*, 2003[54]
DISCO	diagnostic	child/adult	interview + clinician rated	AD/AspD	see text	Wing *et al.*, 2002[55]
ADOS	diagnostic	child/adult	semi-structured interactive session	AD/AspD	see text	Lord *et al.*, 1994[56]

Note: ABC = Autism Behavior Checklist; AD = autism disorder; ADI = Autism Diagnostic Interview–Revised; ADOS = Autism Diagnostic Observation Schedule; AQ = Autism Quotient; ASDI = Asperger Syndrome Diagnostic Interview; ASDS = Asperger Syndrome Diagnostic Scale; AspD = Asperger's disorder; ASQ = Autism Screening Questionnaire; CARS = Childhood Autism Rating Scale; CAST = Childhood Autism Screening Test; M-CHAT = Checklist for Autism in Toddlers; CSBS-DP-IT-Checklist = Communication and Symbolic Behavior Scales Developmental Profile Infant-Toddler Checklist; DISCO = Diagnostic Interview for Social and Communication Disorders; GADS = Gilliam Asperger's Disorder Scale; Parent = primary caregiver; SRS = Social Responsiveness Scales.

[a]Note that these instruments may need to be revised to provide evidence of validity for DSM-5 ASD and supplement but DO NOT REPLACE clinical diagnosis.

over-reactivity.[62] Overt clinical depression is sometimes observed and this may be particularly true for adolescents with Asperger's disorder.[15] Case reports and case series have suggested possible associations with bipolar disorders and tics and Tourette's syndrome. Bullying involvement, including victimization and perpetration, occurs more frequently in general educational settings.[63]

Attentional difficulties also are frequent in autism, reflecting cognitive, language, and social problems.[64] The historical prohibition on making an additional diagnosis of attention-deficit/hyperactivity disorder in those with ASD has been removed in the *DSM-5*. Notably, a subset of children with ASD with elevated scores for hyperactivity showed a 49% response rate in a large randomized controlled trial of methylphenidate treatment.[64]

EVIDENCE BASE FOR PRACTICE PARAMETERS

In this Parameter, recommendations for best assessment and treatment practices are stated in accordance with the strength of the underlying empirical and/or clinical support.

- Clinical standard [CS] is applied to recommendations that are based on rigorous empirical evidence (e.g., meta-analyses, systematic reviews, individual randomized controlled trials) and/or overwhelming clinical consensus.
- Clinical guideline [CG] is applied to recommendations that are based on strong empirical evidence (e.g., nonrandomized controlled trials, cohort studies, case-control studies) and/or strong clinical consensus.
- Clinical option [OP] is applied to recommendations that are based on emerging empirical evidence (e.g., uncontrolled trials or case series/reports) or clinical opinion but lack strong empirical evidence and/or strong clinical consensus.
- Not endorsed [NE] is applied to practices that are known to be ineffective or contraindicated.

The strength of the empirical evidence is rated in descending order as follows:

- [rct] Randomized controlled trial is applied to studies in which subjects are randomly assigned to at least 2 treatment conditions.
- [ct] Controlled trial is applied to studies in which subjects are nonrandomly assigned to at least 2 treatment conditions.
- [ut] Uncontrolled trial is applied to studies in which subjects are assigned to 1 treatment condition.
- [cs] Case series/report is applied to a case series or a case report.

ASSESSMENT

Recommendation 1. The developmental assessment of young children and the psychiatric assessment of all children should routinely include questions about ASD symptomatology [CS].

Screening should include inquiries about the core symptoms of ASD, including social relatedness and repetitive or unusual behaviors. Screening instruments have been developed that may be helpful to the clinician. Some of these instruments are completed by clinicians and others by primary caregivers (Table 1).[43-56] Screening is applicable to young children and to infants, when the diagnosis may first be considered. In some instances, screening may be relevant to older children, e.g., those who are more intellectually able and whose social disability is therefore more likely to be detected later.

Recommendation 2. If the screening indicates significant ASD symptomatology, a thorough diagnostic evaluation should be performed to determine the presence of ASD [CS].

Currently, biological diagnostic markers are not available and diagnosis rests on careful examination of the child. A standard psychiatric assessment should be followed,[65] including interviews with the child and family and a review of past records and historical information. The history and examination should be conducted with careful consideration of *DSM-5* diagnostic criteria. Although the *DSM-5* criteria are intended to be independent of age and intellect, the diagnosis of autism in infants and very young children is more challenging, and some features (e.g., stereotyped movements) may develop later.[5] Systematic attention to the areas relevant to differential diagnosis is essential. Information on the nature of changes over the course of development, e.g., in response to intervention, is helpful. The history should include a review of past and current educational and behavioral interventions and information regarding family history and relevant psychosocial issues. Consideration of possible comorbid diagnoses is an important focus of assessment.

Observation of the child should focus on broad areas of social interaction and restricted, repetitive behaviors. The child's age and developmental level may dictate some modification in assessment procedures. Clinicians should be sensitive to ethnic, cultural, or socioeconomic factors that may affect assessment.

Various instruments for the assessment of ASD have been developed (Table 1[43-56], see Coonrod and Stone[66] for a review). As a practical matter, all these instruments vary in their usefulness for usual clinical practice. Some require specific training. The use of such instruments supplements, but does not replace, informed clinical judgment.[3]

Recommendation 3. Clinicians should coordinate an appropriate multidisciplinary assessment of children with ASD [CS].

All children with ASD should have a medical assessment, which typically includes physical examination, a hearing screen, a Wood's lamp examination for signs of tuberous sclerosis, and genetic testing, which may include G-banded karyotype, fragile X testing, or chromosomal microarray. In a community sample of children with ASD, diagnostic yields were 2.5% for karyotype testing, 0.57% for fragile X testing, and 24% for chromosomal microarray.[67] Chromosomal microarray has been recommended by medical geneticists as the standard of care for the initial evaluation of children with developmental disabilities and/or ASDs.[68] These tests currently detect known abnormalities clearly associated with increased rates of ASD (e.g., 15q11-13 maternal duplications and duplications and deletions of chromosome 16p11.2) and genetic variations of uncertain significance. Recent data from a study of families with only a single affected child have shown that lower IQ is not a strong predictor of a positive chromosomal finding.[69] Any abnormal or indeterminate result from such a study warrants referral for further genetic evaluation and counseling. The yield of genetic testing in the presence of clinical suspicion is currently in the range of at least one third of cases.[70]

Unusual features in the child (e.g., history of regression, dysmorphology, staring spells, family history) should prompt additional evaluations. The list of potential organic etiologies is large but falls into the categories of infectious (e.g., encephalitis or meningitis), endocrinologic (e.g., hypothyroidism), metabolic (e.g., homocystinuria), traumatic (e.g., head injury), toxic (e.g., fetal alcohol syndrome),[4] or genetic (e.g., chromosomal abnormality). Certain developmental disorders, most notably Landau-Kleffner syndrome, also should be ruled out. In this condition, a highly distinctive EEG abnormality is present and associated with development of a marked aphasia.[71] Genetic or neurologic consultation, neuroimaging, EEG, and additional laboratory tests should be obtained when relevant, based on examination or history (e.g., testing for the *MeCP2* gene in cases of possible Rett's disorder).[72]

Psychological assessment, including measurements of cognitive ability and adaptive skills, is indicated for treatment planning and helps to frame observed social-communication difficulties relative to overall development. The results of standard tests of intelligence may show considerable scatter. Unusual islets of ability ("splinter skills") may be present. For children with autism, these sometimes take the form of unusual ability ("savant skills"), e.g., the ability to produce intricate drawings or engage in calendar calculations. For higher functioning children, areas of special interest are often present and the single-minded pursuit of these interests may interfere with the child's ability to learn. Psychological tests clarify areas of strength and weakness useful in designing intervention programs and may need to include instruments valid for a nonverbal population.[7]

Communication assessment, including measurements of receptive and expressive vocabulary and language use (particularly social or pragmatic), is helpful for diagnosis and treatment planning.[73] Occupational and physical therapy evaluations may be needed to evaluate sensory and/or motor difficulties.[74] Sleep is an important variable to assess in individuals with ASD.[75] When members of multiple disciplines are involved in assessment, it is optimal that coordination occur among the various professionals.

TREATMENT

Recommendation 4. The clinician should help the family obtain appropriate, evidence-based, and structured educational and behavioral interventions for children with ASD [CS].

Structured educational and behavioral interventions have been shown to be effective for many children with ASD[76] and are associated with better outcome.[8] As summarized in the National Research Council report,[76] the quality of the research literature in this area is variable,

with most studies using group controls or single-subject experimental methods. In general, studies using more rigorous randomized group comparisons are sparse, reflecting difficulties in random assignment and control comparisons. Other problems include lack of attention to subject characterization, generalization of treatment effects, and fidelity of treatment implementation. Despite these problems, various comprehensive treatments approaches have been shown to have efficacy for groups of children, although none of the comprehensive treatment models has clearly emerged as superior.[76]

Behavioral

Behavioral interventions such as Applied Behavioral Analysis (ABA) are informed by basic and empirically supported learning principles.[77] A widely disseminated comprehensive ABA program is Early Intensive Behavioral Intervention for young children, based on the work of Lovaas *et al.*[78] Early Intensive Behavioral Intervention is intensive and highly individualized, with up to 40 hours per week of one-to-one direct teaching, initially using discrete trials to teach simple skills and progressing to more complex skills such as initiating verbal behavior. A meta-analysis found Early Intensive Behavioral Intervention effective for young children but stressed the need for more rigorous research to extend the findings.[79] Behavioral techniques are particularly useful when maladaptive behaviors interfere with the provision of a comprehensive intervention program. In such situations, a functional analysis of the target behavior is performed, in which patterns of reinforcement are identified and then various behavioral techniques are used to promote a desired behavioral alternative. ABA techniques have been repeatedly shown to have efficacy for specific problem behaviors,[80] and ABA has been found to be effective as applied to academic tasks,[81[ut]] adaptive living skills,[82[ut]] communication,[83[ut]] social skills,[84[ut]] and vocational skills.[85[ct]] Because most children with ASD tend to learn tasks in isolation, an explicit focus on generalization is important.[86]

Communication

Communication is a major focus of intervention and typically will be addressed in the child's individualized educational plan in coordination with the speech-language pathologist. Children who do not yet use words can be helped through the use of alternative communication modalities, such as sign language, communication boards, visual supports, picture exchange, and other forms of augmentative communication. There is some evidence for the efficacy of the Picture Exchange Communication System, sign language, activity schedules, and voice output communication aids.[87[rct],88-90] For individuals with fluent speech, the focus should be on pragmatic language skills training. Children and adolescents with fluent speech may, for example, be highly verbal but have severely impaired pragmatic language skills that can be addressed through explicit teaching. Many programs to enhance social reciprocity and pragmatic language skills are currently available (Table 2; see Reichow and Volkmar[91] for an extensive review).[92-103]

Educational

There is consensus that children with ASD need a structured educational approach with explicit teaching.[76] Programs shown to be effective typically involve planned, intensive, individualized intervention with an experienced, interdisciplinary team of providers, and family involvement to ensure generalization of skills. The educational plan should reflect an accurate assessment of the child's strengths and vulnerabilities, with an explicit description of services to be provided, goals and objectives, and procedures for monitoring effectiveness. Although the curricula used vary across programs, they often share goals of enhancing verbal and nonverbal communication, academic skills, and social, motor, and behavioral capabilities. In some instances, particularly for younger children, a parent-education and home component may be important. Development of an appropriate individualized educational plan is central in providing effective service to the child and family. Efficacy has been shown for 2 of the structured educational models, the Early Start Denver Model[104[rct]] and the Treatment and Education of Autism and related Communication handicapped Children program,[105[ct]] but significant challenges remain in disseminating knowledge about effective interventions to educators.

Other Interventions

There is a lack of evidence for most other forms of psychosocial intervention, although cognitive behavioral therapy has shown efficacy for anxiety and anger management in high functioning youth with ASD.[106[rct],107[rct]] Studies of sensory oriented interventions, such as auditory integration

TABLE 2 Methods Available for the Delivery of Social Reciprocity/Pragmatic Language-Oriented Interventions

Developmental Level	Method	Notes	Reference
Infant/preschool (play based)	guided participation	adult coaching and mediation by trained peers	Schuler and Wolfberg, 2002[92]
	Do-Watch-Listen-Say	careful selection of play materials to foster participation; organization of environment to facilitate participation and cooperation	Quill, 2000[93]
	play organizers	neurotypical peers taught to encourage sharing, helping, and praising to facilitate play; some evidence of generalization	Strain *et al.*, 1977[94]
	buddy skills	teaches neurotypical peers to stay with, play with, and talk to their "buddies"; some evidence of improvement in the frequency of social communication that was generalized to other interactions	Goldstein and Wikstrom, 1996[95]
School age	social stories	state a problem and give the child an acceptable response to it; usually focuses on maladaptive behaviors; little evidence of generalization and maintenance	Gray, 2000[96]
	social skills groups	see text	Kamps *et al.*, 1997[97]
	peer network/circle of friends	typical peers taught to initiate and model appropriate social interactions; results have shown improvement in interaction and generalization to new settings	Kamps *et al.*, 1997[97]; Whitaker *et al.*, 1998[98]
Adolescence	peer network/circle of friends	see above	Whitaker *et al.*, 1998[98]; Paul, 2003[99]
	visual schedule/verbal rehearsal	using written and pictorial representations of expected activities and behavior	Klin and Volkmar, 2000[100]; Hodgdon, 1995[101]
	social skills group	see text	Paul, 2003[99]
	social thinking	addresses underlying social cognitive knowledge required for expression of related social skills; promotes teaching the "why" behind socialization	Crooke *et al.*, 2007[102]
	training scripts	scripts are provided that give the opportunity to ask questions in response to others = initiation of conversation	Klin and Volkmar, 2000[103]

training, sensory integration therapy, and touch therapy/massage, have contained methodologic flaws and have yet to show replicable improvements.[108,109] There is also limited evidence thus far for what are usually termed developmental, social-pragmatic models of intervention, such as Developmental-Individual Difference-Relationship Based/Floortime, Relationship Development Intervention, Social Communication Emotional Regulation and Transactional Support, and Play and Language for Autistic Youths, which generally use naturalistic techniques in the child's community setting to develop social communication abilities. Children with ASD are psychiatrically hospitalized at substantially higher rates than the non-ASD child population.[110] The efficacy of this intervention is unknown, although there

TABLE 3 Randomized Controlled Trials of Psychotropic Medications in Children and Adolescents With Autism Spectrum Disorder (ASD)

Agent	Study	Target Symptoms	Dose	Demographics	Significant Side Effects	Primary Outcome(s)
α_2 Agonists						
Clonidine	Jaselskis *et al.*, 1992[116]	hyperactivity, irritability, inappropriate speech, stereotypy	0.15-0.20 mg divided 3×/d	8 children 5-13 y old	hypotension, drowsiness	statistically and clinically relevant decrease in ABC Irritability subscale
Guanfacine	Handen *et al.*, 2008[117]	hyperactivity, inattention	1-3 mg divided 3×/d	7 children with ASD 5-9 y old	drowsiness, irritability	45% with >50% decrease in ABC Hyperactivity subscale
Antipsychotics						
Aripiprazole	[b]Marcus *et al.*, 2009[118]	irritability, hyperactivity, stereotypy, social withdrawal, inappropriate speech	5, 10, or 15 mg/d fixed dose	218 children 6-17 y old	somnolence, weight gain, drooling, tremor, fatigue, vomiting	56% positive response[a] for aripiprazole 5 mg vs. 35% on placebo; significant improvement in Irritability, Hyperactivity, and Stereotypy subscales
	[b]Owen *et al.*, 2009[119]	irritability, hyperactivity, stereotypy, social withdrawal inappropriate speech	5-15 mg/d flexibly dosed	98 children 6-17 y old	somnolence, weight gain, drooling, tremor, fatigue, vomiting	52% positive response[a] for aripiprazole vs. 14% on placebo; significant improvement in Irritability, Hyperactivity, and Stereotypy subscales
Haloperidol	Anderson *et al.*, 1984[120]	multiple behavioral symptoms, global functioning	0.5-4 mg/d	40 children 2-7 y old	sedation, irritability, extrapyramidal symptoms (>25%)	behavioral symptoms improved with significant decrease in 8 of 14 items of CPRS
	Anderson *et al.*, 1989[121]	multiple behavioral symptoms, global functioning	0.25-4 mg/d	45 children 2-7 y old	sedation, extrapyramidal symptoms	behavioral symptoms improved with significant decrease in 7 of 14 items of CPRS
Olanzapine	[b]Hollander *et al.*, 2006[122]	global functioning, aggression, compulsions, irritability	7.5-12.5 mg/d	11 children 6-14 y old	weight gain, sedation	50% of those on olanzapine much or very much improved in global functioning vs. 20% on placebo
Risperidone	RUPP, 2002[123]	irritability, hyperactivity, stereotypy, social withdrawal, inappropriate speech	0.5-3.5 mg/d	101 children 5-17 y old	weight gain, increased appetite, fatigue, drowsiness, drooling, dizziness	69% had positive response[a] on risperidone vs. 12% positive response[a] on placebo; significant positive findings for hyperactivity and stereotypy

TABLE 3 Continued

Agent	Study	Target Symptoms	Dose	Demographics	Significant Side Effects	Primary Outcome(s)
	[b]Shea *et al.*, 2004[124]	irritability, hyperactivity, stereotypy, social withdrawal, inappropriate speech	0.02-0.06 mg/kg/d	79 children 5-12 y old	weight gain, somnolence,	64% improvement in ABC Irritability subscale on risperidone vs. 31% improvement on placebo; significant positive finding for hyperactivity
	McDougle *et al.*, 2005[125]	social and communication impairment, repetitive behavior and stereotypy	0.5-3.5 mg/d	101 children 5-17 y old	weight gain, increased appetite, fatigue, drowsiness, drooling, dizziness	significant response[c] for repetitive behavior and stereotypy on risperidone
Risperidone vs. haloperidol	[b]Miral *et al.*, 2008[126]	behavior, social, sensory, language	0.01-0.08 mg/kg/d	30 children 8-18 y old	EPS, weight gain, gynecomastia	risperidone reported superior to haloperidol only on ABC total score, no subscales reported
Mood stabilizers						
Valproic acid	Hellings *et al.*, 2005[127]	irritability	20 mg/kg/d, average level 75-78	30 subjects 6-20 y old	increased appetite, skin rash	no significant difference for ABC Irritability subscale
	[b]Hollander *et al.*, 2005[128]	repetitive behavior	500-1,500 mg/d	12 children 5-17 y old, 1 adult 40 y old	irritability, aggression	statistically significant decrease in repetitive behavior on CY-BOCS
	Hollander *et al.*, 2010[129]	global irritability	dosed to mean level of 89.8 μg/mL	27 children 5-17 y old	skin rash, irritability	62.5% positive response for irritability on CGI on divalproex vs. 9.09% on placebo
Lamotrigine	[b]Belsito *et al.*, 2001[130]	irritability, social behavior	5 mg/kg/d	28 children 3-11 y old	insomnia, hyperactivity	no significant difference in irritability or social behavior on multiple instruments
Levetiracetam	[b]Wasserman *et al.*, 2006[131]	irritability, global functioning	20-30 mg/kg/d	20 children 5-17 y old	aggression	no significant difference in global functioning or irritability
Norepinephrine reuptake inhibitors						
Atomoxetine HCl	[b]Harfterkamp *et al.*, 2012[132]	hyperactivity, inattention	1.2 mg/kg/d	97 children 6-17 y old	nausea, anorexia, fatigue, early wakening	significant difference in the ADHD-RS for active treatment group; no difference in CGI

TABLE 3 Continued

Agent	Study	Target Symptoms	Dose	Demographics	Significant Side Effects	Primary Outcome(s)
	[b]Arnold *et al.*, 2006[133]	hyperactivity, inattention	20-100 mg divided 2×, mean 44 mg/d	16 children 5-15 y old	upper GI symptoms, fatigue, racing heart	57% positive response[a] for parent-rated ABC Hyperactivity subscale vs. 25% on placebo
Serotonin reuptake inhibitors						
Citalopram	King *et al.*, 2009[134]	repetitive behavior	2.5-20 mg/d, mean 16 mg/d	149 children 5-17 y old	hyperactivity, insomnia, inattention, impulsivity, diarrhea, stereotypy	no significant difference in repetitive behavior on CGI-I and CY-BOCS PDD
Fluoxetine	Hollander *et al.*, 2005[135]	repetitive behavior	2.4-20 mg/d, mean 9.9 mg/d	39 children 5-17 y old	none significant	statistically significant decrease in repetitive behavior on CY-BOCS Compulsions scale
Clomipramine	Gordon *et al.*, 1993[136]	stereotypy, repetitive behavior, compulsions	25-250 mg/d, mean 152 mg/d	12 children 6-18 y old	insomnia, constipation, twitching, tremors	decrease in repetitive behavior on CPRS
	Remington *et al.*, 2001[137]	stereotypy, irritability, hyperactivity	100-150 mg/d, mean 128.4 mg/d	31 subjects <20 y old	lethargy, tremors, tachycardia, insomnia, diaphoresis, nausea	no significant difference in stereotypy, irritability, or hyperactivity for clomipramine on ABC
Stimulants						
Methylphenidate	RUPP, 2005[138]	hyperactivity	7.5-50 mg/d divided 3×/d	58 children 5-14 y old	decreased appetite, insomnia, irritability, emotionality	49% positive responders[a] for hyperactivity vs. 15.5% on placebo
	Pearson *et al.*, 2013[139]	hyperactivity, inattention	10-40 mg each morning, methylphenidate extended release	24 children 7-12 y old	decreased appetite, insomnia	significant decrease in hyperactivity and inattention on multiple teacher and parent measurements
	Handen *et al.*, 2000[140]	hyperactivity	0.3-0.6 mg/kg/dose, 2-3×/d	13 children 5-11 y old	social withdrawal, irritability	8 of 13 children with >50% decrease in hyperactivity on Teacher Conners Hyperactivity subscale
	Quintana *et al.*, 1995[141]	hyperactivity	10-20 mg 2×/d	10 children 7-11 y old	irritability, anorexia, insomnia	decrease in ABC Hyperactivity subscale by 8 points over placebo
Miscellaneous						
Amantadine	[b]King *et al.*, 2001[142]	hyperactivity, irritability	2.5-5.0 mg/kg/d	39 children 5-19 y old	insomnia	no statistical difference in parent ABC Hyperactivity or Irritability subscales, statistical improvement in clinician Hyperactivity and Inappropriate Speech subscales

TABLE 3 Continued

Agent	Study	Target Symptoms	Dose	Demographics	Significant Side Effects	Primary Outcome(s)
Cyproheptadine (in combination with haloperidol)	Akhondzadeh *et al.*, 2004[143]	ABC total score, CARS	Titrated up to 0.2 mg/kg/d	40 children 3-11 y old	none significant, trend toward increased appetite	statistically significant difference in ABC total score and CARS diagnostic screening tool, with unknown clinical significance
Donepezil	Chez *et al.*, 2003[144]	"autistic behavior," expressive-receptive communication	1.25-2.5 mg/d	43 children 2-10 y old	diarrhea, stomach cramping, irritability	"autistic behavior" statistically, improved on CARS diagnostic screening tool with unknown clinical significance
Naltrexone	Willemsen-Swinkels *et al.*, 1995[145]	"social behavior," irritability	single 40-mg dose	20 children 3-7 y old	sedation, increased stereotypy	no effect on social behavior; significant decrease on ABC Irritability subscale vs. placebo
	[b]Kolmen *et al.*, 1995[146]	hyperactivity, communication initiation	1 mg/kg/d	13 children 3-8 y old	transient sedation	no significant difference in communication initiation
	[b]Feldman *et al.*, 1999[147]	communication	1 mg/kg/d	24 children, 3-8 y old	transient sedation	no significant difference in multiple communication measurements
	Campbell *et al.*, 1993[148]	CGI, CPRS, discriminant learning, hyperactivity	0.5-1 mg/kg/d	18 children 3-8 y old	increased aggression and stereotypy	no significant difference on CGI or CPRS or discriminant learning; positive trend for hyperactivity
	Campbell *et al.*, 1990[149]	hyperactivity, discriminant learning, self-injurious behavior	0.5-1 mg/kg/d	41 children 3-8 y old	none significant	significantly decreased hyperactivity; no effect on discriminant learning; positive trend for self-injurious behavior
Pentoxifylline (in combination with risperidone)	Akhondzadeh *et al.*, 2010[150]	irritability, hyperactivity, stereotypy, social withdrawal, inappropriate speech	200-600 mg/d	40 children 4-12 y old	sedation, GI effects, increased appetite	significant improvement on ABC Irritability and Social Withdrawal subscales

Note: ABC = Autism Behavior Checklist; ADHD-RS = Attention-Deficit/Hyperactivity Disorder Rating Scale; CY-BOCS = Children's Yale-Brown Obsessive Compulsive Scale; CARS = Childhood Autism Rating Scale; CPRS = Children's Psychiatric Rating Scale; EPS = extrapyramidal side effects; GI = gastrointestinal; PDD = pervasive developmental disorder; RUPP = Research Units on Pediatric Psychopharmacology.

[a]A positive response in this study was defined as a >25% reduction in the ABC subscale and a Much Improved or Very Much Improved rating on the Clinical Global Impression–Global Improvement (CGH).

[b]Study identified as funded by pharmaceutical industry.

[c]A positive response in this study was defined as a greater than 25% decrease in ABC (CY-BOCS) compulsions score and a much improved or very much improved rating on the CGH.

is preliminary evidence for the efficacy of hospital psychiatry units that specialize in the population.[111]

Recommendation 5. Pharmacotherapy may be offered to children with ASD when there is a specific target symptom or comorbid condition [CG].

Pharmacologic interventions may increase the ability of persons with ASD to profit from educational and other interventions and to remain in less restrictive environments through the management of severe and challenging behaviors. Frequent targets for pharmacologic intervention include associated comorbid conditions (e.g., anxiety, depression) and other features, such as aggression, self-injurious behavior, hyperactivity, inattention, compulsive-like behaviors, repetitive or stereotypic behaviors, and sleep disturbances. As with other children and adolescents, various considerations should inform pharmacologic treatment.[112] Risperidone[113[rct]] and aripiprazole[114[rct]] have been approved by the Food and Drug Administration for the treatment of irritability, consisting primarily of physical aggression and severe tantrum behavior, associated with autism. There is a growing body of controlled evidence for pharmacologic intervention,[115] and a summary of randomized controlled trials of medication in children with ASD is included (Table 3).[116-150] Combining medication with parent training is moderately more efficacious than medication alone for decreasing serious behavioral disturbance and modestly more efficacious for adaptive functioning.[151[rct],152[rct]] Individuals with ASD may be nonverbal, so treatment response is often judged by caregiver report and observation of specific behaviors. Although this may help document the effectiveness of the selected medication, one must remember that an overall goal of treatment is to facilitate the child's adjustment and engagement with educational intervention. Several objective rating scales also are available to help monitor treatment response.[153]

TABLE 4 Resources for Parents

ASPEN TM, Inc. (Asperger Syndrome Education Network) (http://www.aspennj.org)	A regional nonprofit organization providing families and those individuals affected with Asperger syndrome and related disorders with information, support, and advocacy.
Autism Society of America (http://www.autism-society.org)	The mission of the Autism Society of America is to promote lifelong access and opportunities for persons within the autism spectrum and their families to be fully included, participating members of their communities through advocacy, public awareness, education, and research related to autism.
Autism Speaks (http://www.autismspeaks.org)	Autism Speaks is an autism science and advocacy organization dedicated to funding research into the causes, prevention, treatments, and a cure for autism; increasing awareness of autism spectrum disorders; and advocating for the needs of individuals with autism and their families.
Division TEACCH (Treatment and Education of Autism and related Communication handicapped Children, University of North Carolina at Chapel Hill) (www.teacch.com)	The TEACCH Web site includes information about their program, educational and communication approaches to teaching individuals with autism, their research and training opportunities, and information and resources on autism.
LDAA (Learning Disabilities Association of America) (http://www.ldanatl.org)	The LDAA site includes information and resources on many learning disabilities, including learning disabilities involving a significant social component, such as autism and Asperger syndrome.
OASIS (Online Asperger Syndrome Information and Support) (http://www.asperger.org)	General information on Asperger syndrome and related disorders, including resources and materials, announcements of major pertinent events and publications, and being the major "intersection" for communication among parents, clinicians, educators, and individuals with social disabilities.
Yale Child Study Center (www.autism.fm)	Information on autism, Asperger syndrome, and related disorders, lists of resources organized by state, and parent support organizations and advocacy agencies.

Recommendation 6. The clinician should maintain an active role in long-term treatment planning and family support and support of the individual [CG].

Children's and families' need for help and support will change over time. The clinician should develop a long-term collaboration with the family and realize that service utilization may be sporadic. For very young children, issues of diagnosis and identification of treatment programs often will be most important. For school-age children, psychopharmacologic and behavioral issues typically become more prominent. For adolescents, vocational and prevocational training and thoughtful planning for independence/self-sufficiency is important. As part of this long-term engagement, parents and siblings of children with ASD will need support (Table 4). Although raising a child with autism presents major challenges, rates of parental separation and divorce are not higher among parents of children with ASD than those with non-ASD children.[154]

Recommendation 7. Clinicians should specifically inquire about the use of alternative/complementary treatments and be prepared to discuss their risk and potential benefits [CS].

Although most alternative or complementary treatment approaches have very limited empirical support for their use in children with ASD, they are commonly pursued by families.[155] It is important that the clinician be able to discuss these treatments with parents, recognizing the motivation for parents to seek all possible treatments. In most instances, these treatments have little or no proved benefit but also have little risk.[7] In a few instances, the treatment has been repeatedly shown not to work (e.g., intravenous infusion of secretin[156] and oral vitamin B6 and magnesium[157[rct]]), or randomized controlled evidence does not support its use (e.g., the gluten-free, casein-free diet,[158] ω-3 fatty acids,[159] and oral human immunoglobulin).[160[rct]] Some treatments have greater potential risk to the child directly (e.g., mortality and morbidity associated with chelation[161[cs]]) or from side effects owing to contaminants in "natural" compounds or indirectly (e.g., by diverting financial or psychosocial resources). For a detailed review of alternative treatments, see Jacobson *et al.*[162] and Levy and Hyman.[163] Although more controlled studies of these treatments are needed, it is important that the family be able to voice their questions to health care providers. Families may be guided to the growing body of work on evidence-based treatments in autism.[164]

PARAMETER LIMITATIONS

AACAP Practice Parameters are developed to assist clinicians in psychiatric decision making. These Parameters are not intended to define the sole standard of care. As such, the Parameters should not be deemed inclusive of all proper methods of care or exclusive of other methods of care directed at obtaining the desired results. The ultimate judgment regarding the care of a particular patient must be made by the clinician in light of all of the circumstances presented by the patient and his or her family, the diagnostic and treatment options available, and available resources. &

This Parameter was developed by Fred Volkmar, MD, Matthew Siegel, MD, Marc Woodbury-Smith, MD, Bryan King, MD, James McCracken, MD, Matthew State, MD, PhD, and the American Academy of Child and Adolescent Psychiatry (AACAP) Committee on Quality Issues (CQI): William Bernet, MD, Oscar G. Bukstein, MD, MPH, and Heather J. Walter, MD, MPH, co-chairs; and Christopher Bellonci, MD, R. Scott Benson, MD, Regina Bussing, MD, Allan Chrisman, MD, Tiffany R. Farchione, MD, John Hamilton, MD, Munya Hayek, MD, Helene Keable, MD, Joan Kinlan, MD, Nicole Quiterio, MD, Carol Rockhill, MD, Ulrich Schoettle, MD, Matthew Siegel, MD, and Saundra Stock, MD.

The AACAP Practice Parameters are developed by the AACAP CQI in accordance with American Medical Association policy. Parameter development is an iterative process between the primary author(s), the CQI, topic experts, and representatives from multiple constituent groups, including the AACAP membership, relevant AACAP committees, the AACAP Assembly of Regional Organizations, and the AACAP Council. Details of the Parameter development process can be accessed on the AACAP Web site. Responsibility for Parameter content and review rests with the author(s), the CQI, the CQI Consensus Group, and the AACAP Council.

The AACAP develops patient-oriented and clinician-oriented Practice Parameters. Patient-oriented Parameters provide recommendations to guide clinicians toward best assessment and treatment practices. Recommendations are based on the critical appraisal of empirical evidence (when available) and clinical consensus (when not) and are graded according to the strength of the empirical and clinical support. Clinician-oriented Parameters provide clinicians with the information (stated as principles) needed to develop practice-based skills. Although empirical evidence may be available to support certain principles, principles are based primarily on clinical consensus. This Parameter is a patient-oriented Parameter.

The primary intended audience for the AACAP Practice Parameters is child and adolescent psychiatrists; however, the information contained therein also may be useful for other mental health clinicians.

The authors acknowledge the following experts for their contributions to this Parameter: Andrés Martin, MD, Schuyler Henderson, MD, Rhea Paul, PhD, Joaquin Fuentes, MD, Christopher McDougle, MD, Ami Klin, PhD, and Connie Zajicek, MD.

Kristin Kroeger Ptakowski and Jennifer Medicus served as the AACAP staff liaisons for the CQI.

This Practice Parameter was reviewed at the Member Forum at the AACAP annual meeting in October 2006.

From March to June 2012, this Parameter was reviewed by a consensus group convened by the CQI. Consensus group members

and their constituent groups were Oscar G. Bukstein, MD, co-chair; R. Scott Benson, MD, and John Hamilton, MD (CQI); Doug Novins, MD, and Christopher Thomas, MD (topic experts); Bryan King, MD (AACAP Autism and Intellectual Disability Committee); Melissa Del-Bello, MD (AACAP Research Committee); John Rose, MD, and Syed Naqvi, MD (AACAP Assembly of Regional Organizations); and Louis Kraus, MD, and Tami Benton, MD (AACAP Council).

This Practice Parameter was approved by the AACAP Council on July 8, 2013.

This Practice Parameter is available on the Internet (http://www.aacap.org).

Disclosures: Fred Volkmar, MD, receives or has received research funding from the National Institute of Child Health and Human Development and the National Institute of Mental Health and has intellectual property with John Wiley & Sons, Inc., Guilford Publications, Inc, and Springer. Matthew Siegel, MD, has no financial conflicts of interest to disclose. Marc Woodbury-Smith, MD, has no financial conflicts of interest to disclose. Bryan King, MD, has or has received research funding from the National Institutes of Health (NIH), Seaside Therapeutics, and Health Resources and Services Administration and serves or has served as an advisor/consultant with the U.S. Department of Justice. James McCracken, MD, has or has received research funding from Seaside Therapeutics and Bristol-Myers Squibb, serves or has served as an advisor/consultant to BioMarin Pharmaceuticals, Inc., and receives or has received honoraria as a speaker for Veritas, Discovery Channel Health CME, and CME Outfitters, LLC. Matthew State, MD, has or has received research funding from the NIH and Howard Hughes Medical Institute and has an exclusive license agreement with Athena Diagnostics. Oscar Bukstein, MD, MPH, co-chair, has served as a consultant for Ezra Innovations and for PRIME CME. He receives royalties from Routledge Press. Heather Walter, MD, MPH, and William Bernet, MD, co-chairs, have no financial relationships to disclose. Disclosures of potential conflicts of interest for all other individuals named above are provided on the AACAP Web site on the Practice Parameters page.

Correspondence to the AACAP Communications Department, 3615 Wisconsin Avenue, NW, Washington, D.C. 20016.

0890-8567/$36.00/©2014 American Academy of Child and Adolescent Psychiatry

http://dx.doi.org/10.1016/j.jaac.2013.10.013

REFERENCES

1. American Academy of Child and Adolescent Psychiatry. Practice parameters for the assessment and treatment of children, adolescents, and adults with autism and other pervasive developmental disorders. J Am Acad Child Adolesc Psychiatry. 1999;38 (suppl):32S-54S.
2. Kanner L. Autistic disturbances of affective contact. Nervous Child. 1943;2:217-250.
3. Volkmar FR, Klin A. Issues in the classification of autism and related conditions. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005:5-41.
4. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorder, 4th ed, text rev. Washington, DC: American Psychiatric Press; 2000.
5. Chawarska K, Klin A, Volkmar FR, eds. Autism Spectrum Disorders in Infants and Toddlers: Diagnosis, Assessment, and Treatment. New York: Guilford Press; 2008.
6. Loveland KA, Tunali-Kotoski B. The school age child with autism. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 2nd ed. New York: Wiley; 1997:283-308.
7. Volkmar FR, Wiesner LA. A Practical Guide to Autism: What Every Parent, Family Member, and Teacher Needs to Know. Hoboken, NJ: John Wiley; 2009.
8. Howlin P. Outcomes in autism spectrum disorders. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005.
9. Amir RE, Van den Veyver IB, Wan M, *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet. 1999;23:185-188.
10. Van Acker R, Loncola JA, Van Acker EY. Rett's syndrome: a pervasive developmental disorder. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005:126-164.
11. Villard L, Kpebe A, Cardoso C, *et al.* Two affected boys in a Rett syndrome family: clinical and molecular findings. Neurology. 2000;55:1188-1193.
12. Clayton-Smith J, Watson P, Ramsden S, *et al.* Somatic mutation in MECP-2 as a nonfatal neurodevelopmental disorder in males. Lancet. 2000;356:830-832.
13. Orrico A, Lam C, Galli L, *et al.* MECP2 mutation in male patients with nonspecific X-linked mental retardation. FEBS Lett. 2000; 481:285-288.
14. Volkmar FR, Koenig K, State M. Childhood disintegrative disorder. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorder. 3rd ed. Hoboken, NJ: Wiley; 2005:70-78.
15. Klin A, McPartland J, Volkmar FR. Asperger syndrome. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorder. 3rd ed. Hoboken, NJ: Wiley; 2005:88-125.
16. Towbin KE. Pervasive developmental disorder not otherwise specified. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005:165-200.
17. Lord C, Petkova E, Hus V, *et al.* A multisite study of the clinical diagnosis of different autism spectrum disorders. Arch Gen Psychiatry. 2012;69:306-313.
18. Fombonne E. Epidemiological studies of pervasive developmental disorders. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005.
19. Prevalence of autism spectrum disorders—Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. MMWR CDC Surveill Summ. 2012;61:1-19.
20. Williams K, Glasson EJ, Wray J. Incidence of autism spectrum disorders in children in two Australian states. Med J Aust. 2005; 182:108-111.
21. Ozonoff S, Rogers SJ, Hendren RL, eds. Autism Spectrum Disorders: A Research Review for Practitioners. Washington, D.C: American Psychiatric Publishing; 2003.
22. Mandell DS, Ittenbach RF, Levy SE, Pinto-Martin JA. Disparities in diagnoses received prior to a diagnosis of autism spectrum disorder. J Autism Dev Disord. 2006;37:1795-1802.
23. Volkmar F, Nelson DS. Seizure disorders in autism. J Am Acad Child Adolesc Psychiatry. 1991;29:127-129.
24. Minshew NJ, Sweeney JA, Bauman ML, *et al.* Neurologic aspects of autism. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005:453-472.
25. Bauman ML, Kemper TL. Neuroanatomic observations of the brain in autism: a review and future directions. Int J Dev Neurosci. 2005;23:183-187.
26. Schultz RT, Robbins DL. Functional neuroimaging studies of autism spectrum disorders. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005: 515-533.
27. Wolff JJ, Gu H, Gerig G, *et al.* Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. Am J Psychiatry. 2012;169:589-600.

28. Anderson GM, Hoshino Y. Neurochemical studies of autism. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders, 3rd ed, vol. 1. Hoboken, NJ: Wiley; 2005:453-472.
29. DeStefano F, Thompson WW. MMR vaccine and autism: an update of the scientific evidence. Expert Rev Vaccines. 2004; 3:19-22.
30. Rutter M. Genetic influences and autism. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005: 425-452.
31. Pardo CA, Vargas DL, Zimmerman AW. Immunity, neuroglia and neuroinflammation in autism. Int Rev Psychiatry. 2005;17: 485-495.
32. Ozonoff S, Pennington BF, Rogers SJ. Executive function deficits in high functioning autistic individuals: relationship to theory of mind. J Child Psychol Psychiatry. 1991;32:1081-1105.
33. Happe F, Frith U. The weak coherence account: detail-focused style in autism spectrum disorders. J Autism Dev Disord. 2006; 36:5-25.
34. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a theory of mind? Cognition. 1985;21:37-46.
35. Ozonoff S, Young GS, Carter A, *et al.* Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. Pediatrics. 2011;128:e488-e495.
36. Cheslack-Postava K, Liu K, *et al.* Closely spaced pregnancies are associated with increased odds of autism in California sibling births. Pediatrics. 2011;127:246-253.
37. Croen LA, Najjar DV, Fireman B, Grether JK. Maternal and paternal age and risk of autism spectrum disorders. Arch Pediatrics Adolesc Med. 2007;161:334-340.
38. Johnson S, Hollis C, Kochhar P, *et al.* Autism spectrum disorders in extremely preterm children. Pediatrics. 2010;156:525-531.e522.
39. Veenstra-VanderWeele J, Christina SL, Cook EH. Autism as a paradigmatic complex genetic disorder. Annu Rev Genomics Hum Genet. 2004;5:379-405.
40. State MW. The genetics of child psychiatric disorders: focus on autism and Tourette syndrome. Neuron. 2010;68:254-269.
41. Abrahams BS, Geschwind DH. Advances in autism genetics on the threshold of a new neurobiology. Nat Rev Genet. 2008;9: 341-355.
42. Chawarska K, Volkmar F. Autism in infancy and early childhood. In: Volkmar F, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. New York: Wiley; 2005:223-247.
43. Krug DA, Arick J, Almond P. Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. J Child Psychol Psychiatry. 1980;21:221-229.
44. Schopler E, Reichler RJ, DeVellis RF, *et al.* Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). J Autism Dev Disord. 1980;10:91-103.
45. Robins DL, Fein D, Barton ML, Green JA. The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. J Autism Dev Disord. 2001;31:131-144.
46. Wetherby AM, Brosnan-Maddox S, Peace V, *et al.* Validation of the infant-toddler checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. Autism. 2008;12: 487-511.
47. Berument SK, Rutter M, Lord C, *et al.* Autism screening questionnaire: diagnostic validity. Br J Psychiatry. 1999;175:444-451.
48. Baron-Cohen S, Wheelwright S, Skinner R, *et al.* The Autism Spectrum Quotient (AQ): evidence from Asperger syndrome/ high functioning autism, males and females, scientists and mathematicians. J Autism Dev Disord. 2001;31:5-17.
49. Scott F, Baron-Cohen S, Bolton P, *et al.* The CAST (Childhood Asperger Syndrome Test): preliminary development of UK screen for mainstream primary-school children. Autism. 2002;6:9-31.
50. Myles BS, Bock SJ, Simpson RL. Asperger Syndrome Diagnostic Scale. Austin, TX: PRO-ED; 2000.
51. Gilliam JE. Gilliam Asperger Disorder Scale. Austin, TX: PRO-ED; 2001.
52. Gillberg C, Gillberg C, Rastam M, *et al.* The Asperger Syndrome (and High-Functioning Autism) Diagnostic Interview (ASDI): a preliminary study of a new structured clinical interview. Autism. 2001;5:57-66.
53. Constantino JN, Hudziak JJ, Todd RD. Deficits in reciprocal social behavior in male twins: evidence for a genetically independent domain of psychopathology. J Am Acad Child Adolesc Psychiatry. 2003;42:458-467.
54. Lord C, Rutter M, DiLavore P, *et al.* Autism Diagnostic Observation Schedule. Los Angeles: Western Psychological Services; 2003.
55. Wing L, Leekam SR, Libby SJ, *et al.* The Diagnostic Interview for Social and Communication Disorders: background, inter-rater reliability and clinical use. J Child Psychol Psychiatry. 2002;43: 307-325.
56. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview–revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24:659-685.
57. Luteijn EF, Serra M, Jackson S, *et al.* How unspecified are disorders of children with a pervasive developmental disorder not otherwise specified? A study of social problems in children with PDD-NOS and ADHD. Eur Child Adolesc Psychiatry. 2000;9: 168-179.
58. Roeyers H, Keymculon H, Buysse A. Differentiating attention deficit/hyperactivity disorder from pervasive developmental disorder not otherwise specified. J Learn Disabil. 1998;34: 565-571.
59. Reiss S, Levitan GW, Szyszko J. Emotional disturbance and mental retardation: diagnostic overshadowing. Am J Ment Defic. 1982;86:567-574.
60. Leyfer OT, Folstein SE, Bacalman S, *et al.* Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. J Autism Dev Disord. 2006;36:849-861.
61. Klin A, Pauls D, Schultz R, Volmar F. Three diagnostic approaches to Asperger syndrome: implications for research. J Autism Dev Disord. 2005;35:241-257.
62. Mazefsky C, White SW, Siegel M, *et al.* The role of emotion regulation in autism spectrum disorder. J Am Acad Child Adolesc Psychiatry. 2013;52:679-688.
63. Sterzing PR, Shattuck PT, Narendorf FC, Wagner M, Cooper BP. Bullying involvement and autism spectrum disorders: prevalence and correlates of bullying involvement among adolescents with an autism spectrum disorder. Arch Pediatr Adolesc Med. 2012; 166:1058-1064.
64. Research Units on Pediatric Psychopharmacology Autism Network. A randomized, double-blind, placebo-controlled, crossover trial of methylphenidate in children with hyperactivity associated with pervasive developmental disorders. Arch Gen Psychiatry. 2005;62:1266-1274.
65. American Academy of Child and Adolescent Psychiatry. Practice parameters for the psychiatric assessment of children and adolescents. J Am Acad Child Adolesc Psychiatry. 1997;36(suppl): 4S-20S.
66. Coonrod EE, Stone WL. Screening for autism in young children. In: Volkmar F, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. New York: Wiley; 2005:707-730.
67. McGrew SG, Peters BR, Crittendon JA, Veenstra-Vanderweele J. Diagnostic yield of chromosomal microarray analysis in an autism primary care practice: which guidelines to implement? J Autism Dev Disord. 2012;42:1582-1591.
68. Miller DT, Adam MP, Aradhya S, *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86:749-764.
69. Sanders SJ, Ercan-Sencicek AG, Hus V, *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70:863-885.
70. Moeschler JB, Shevell M; Committee on Genetics. Clinical genetic evaluation of the child with mental retardation or developmental delays. Pediatrics. 2006;117:2304-2316.
71. Camfield P, Camfield C. Epileptic syndromes in childhood: clinical features, outcomes, and treatment. Epilepsia. 2002;43 (suppl 3):27-32.

72. Schaefer GB, Mendelsohn NJ. Genetics evaluation for the etiologic diagnosis of autism spectrum disorders. Genet Med. 2008; 10:4-12.
73. Paul R, Sutherland D. Enhancing early language in children with autism spectrum disorders. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005: 946-976.
74. Baranek GT, Parham LD, Bodfish JW. Sensory and motor features in autism: assessment and intervention. In: Volkmar FR, Klin A, Paul R, Cohen DJ, eds. Handbook of Autism and Pervasive Developmental Disorders. 3rd ed. Hoboken, NJ: Wiley; 2005:88-125.
75. Goldman SE, Richdale AL, Clemons T, Malow BA. Parental sleep concerns in autism spectrum disorders: variations from childhood to adolescence. J Autism Dev Disord. 2012;42:531-538.
76. National Research Council. Educating Children with Autism. Washington, D.C.: National Academy of Sciences Press; 2001.
77. Cooper JO, Heron TA, Heward WL. Applied Behavioral Analysis. Upper Saddle River, NJ: Prentice Hall; 1987.
78. Lovaas OI, Ackerman A, Alexander D, *et al.* Teaching Developmentally Disabled Children: The ME Book. Austin, TX: PRO-ED; 1981.
79. Howlin P, Magiati I, Charman T. Systematic review of early intensive behavioral interventions for children with autism. Am J Intell Dev Disabil. 2009;114:23-41.
80. Campbell JM. Efficacy of behavioral interventions for reducing problem behavior in people with autism: A quantitative synthesis of single-subject research. Res Dev Disabil. 2003;24: 120-138.
81. Koegel LK, Carter CM, Koegel RL. Teaching children with autism self-initiations as a pivotal response. Topics Lang Disord. 2003;23: 134-145.
82. Leblanc LA, Carr JE, Crossett SE, Bennett CM, Detweiler DD. Intensive outpatient behavioral treatment of primary urinary incontinence of children with autism. Focus Autism Other Dev Disabil. 2005;20:98-105.
83. Jones EA, Feeley KM, Takacs J. Teaching spontaneous responses to young children with autism. J Appl Behav Anal. 2007;40: 565-570.
84. Pierce K, Schreibman L. Increasing complex social behaviors in children with autism: effects of peer implemented pivotal response training. J Appl Behav Anal. 1995;28:285-295.
85. Lattimore LP, Parsons MB, Reid DH. Enhancing job-site training of supported workers with autism: a reemphasis on simulation. J Appl Behav Anal. 2006;39:91-102.
86. Foxx R. Applied behavioral analysis of autism: the state of the art. Child Adolesc Psych Clin North Am. 2008;17:821-834.
87. Yoder P, Stone WL. A randomized comparison of the effect of two prelinguistic communication interventions on the acquisition of spoken communication in preschoolers with ASD. J Speech Lang Hear Res. 2006;49:698-711.
88. Beukelman DR, Mirenda P. Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs. Baltimore, MD; Brooks Publishing; 2005.
89. Lequia J, Machalicek W, Ripoli M. Effects of activity schedules on challenging behavior exhibited in children with autism spectrum disorders: a systematic review. Res Autism Spectrum Disord. 2012;6:480-492.
90. Ganz JB, Earles-Vollrath TL, Heath AK, Parker RI, Rispoli MJ, Duran JB. A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. J Autism Dev Disord. 2012;42:60-74.
91. Reichow B, Volkmar FR. Social skills interventions for individuals with autism: evaluation for evidence-based practices within a best evidence synthesis framework. J Autism Dev Disord. 2010;40:149-166.
92. Schuler AL, Wolfberg PJ. Promoting peer socialization and play: the art of scaffolding. In: Prizant B, Wetherby A, eds. Language Issues in Autism and Pervasive Developmental Disorder: A Transactional Developmental Perspective. Baltimore, MD: Paul H. Brookes; 2002.
93. Quill KA. Do-Watch-Listen-Say: Social and Communication Intervention for Children with Autism. Baltimore, MD: Paul H Brookes; 2000.
94. Strain PS, Shores RE, Timm MA. Effects of peer social initiations on the behavior of withdrawn preschool children. J Appl Behav Anal. 1977;10:289-298.
95. Goldstein H, Wickstrom S. Peer intervention effects on communicative interaction among handicapped and non-handicapped preschoolers. J Appl Behav Anal. 1996;19:209-214.
96. Gray C. The New Social Story Book. Arlington, TX: Future Horizons; 2000.
97. Kamps DM, Potucek J, Lopez AG, Kravits T, Kemmerer K. The use of peer networks across multiple settings to improve social interaction for students with autism. J Behav Educ. 1997;7: 335-357.
98. Whitaker P, Barratt P, Joy H, *et al.* Children with autism and peer group support: using circles of friends. Br J Spec Educ. 1998; 25:60-64.
99. Paul R. Promoting social communication in high functioning individuals with autistic spectrum disorders. Child Adolesc Psychiatr Clin North Am. 2003;12:87-106.
100. Klin A, Volkmar FR, eds. Treatment and Intervention Guidelines for Individuals with Asperger Syndrome. New York: Guilford Press; 2000:340-366.
101. Hodgdon LA. Visual Strategies for Improving Communication: Practical Supports for School and Home. Troy, MI: QuickRoberts Publishing; 1995.
102. Crooke PJ, Hendrix RE, Rachman JY. Brief report: measuring the effectiveness of teaching social thinking to children with Asperger syndrome (AS) and high functioning autism (HFA). J Autism Dev Disord. 2007;38:581-591.
103. Klin A, Volkmar FR, eds. Treatment and Intervention Guidelines for Individuals with Asperger Syndrome. New York: Guilford Press; 2000:340-366.
104. Dawson G, Rogers S, Munson J, *et al.* Randomized, controlled trial of an intervention for toddlers with autism: the early start Denver model. Pediatrics. 2010;125:e17-e23.
105. Ozonoff S, Cathcart K. Effectiveness of a home program intervention for young children with autism. J Autism Dev Disord. 1998;28:25-32.
106. Wood JJ, Drahota A, Sze K, Har K, Chiu A, Langer DA. Cognitive behavioral therapy for anxiety in children with autism spectrum disorders: a randomized controlled trial. J Child Psychol Psychiatry. 2009;50:224-234.
107. Sofronoff K, Attwood T, Hinton S, *et al.* A randomized controlled trial of a cognitive behavioral intervention for anger management in children diagnosed with Asperger syndrome. J Autism Dev Disord. 2007;37:1203-1214.
108. Leong HM, Carter M. Research on the efficacy of sensory integration therapy: past, present and future. Australas J Spec Educ. 2008;32:83-89.
109. Sinha Y, Silove N, Hayen A, Williams K. Auditory integration training and other sound therapies for autism spectrum disorders (ASD). Cochrane Database Syst Rev. 2011;12:CD003681.
110. Croen LA, Najjar DV, Ray T. A comparison of health care utilization and costs of children with and without autism spectrum disorders in a large group model health plan. Pediatrics. 2006; 118:e1203.
111. Siegel M, Gabriels R. Psychiatric hospital treatment for children with autism and serious behavioral disturbance. Child Psychiatry Clin N Am. 2014;23:125-142.
112. American Academy of Child and Adolescent Psychiatry. Practice parameter on the use of psychotropic medication in children and adolescents. J Am Acad Child Adolesc Psychiatry. 2009;48: 961-973.
113. McDougle C, Scahill L, Aman M, *et al.* Risperidone for the core symptom domains of autism: results from the study by the Autism Network of the Research Units on Pediatric Psychopharmacology. Am J Psychiatry. 2005;162:1142-1148.
114. Owen R, Sikich L, Marcus RN, *et al.* Aripiprazole in the treatment of irritability in children and adolescents with autistic disorder. Pediatrics. 2009;124:1533-1540.
115. Siegel M, Beaulieu A. Psychotropic medications in children and adolescents with autism spectrum disorders: a systematic review and synthesis for evidence-based practice. J Autism Dev Disord. 2012;42:1592-1605.
116. Jaselskis CA, Cook EH, Fletcher KE. Clonidine treatment of hyperactive and impulsive children with autistic disorder. J Clin Psychopharmacol. 1992;12:322-327.

117. Handen B, Sahl R, Harden A. Guanfacine in children with autism and/or intellectual disabilities. J Dev Behav Pediatr. 2008;29: 303-308.
118. Marcus R, Owen R, Kamen L, *et al.* A placebo-controlled, fixed-dose study of aripiprazole in children and adolescents with irritability associated with autistic disorder. J Am Acad Child Adolesc Psychiatry. 2009;48:1110-1119.
119. Owen R, Sikich L, Marcus RN, *et al.* Aripiprazole in the treatment of irritability in children and adolescents with autistic disorder. Pediatrics. 2009;124:1533-1540.
120. Anderson LT, Campbell M, Grega DM, *et al.* Haloperidol in infantile autism: effects on learning and behavioral symptoms. Am J Psychiatry. 1984;141:195-202.
121. Anderson LT, Campbell M, Adams P, *et al.* The effects of haloperidol on discrimination learning and behavioral symptoms in autistic children. J Autism Dev Disord. 1989;19:227-239.
122. Hollander E, Wasserman S, Swanson EN, *et al.* A double-blind placebo-controlled pilot study of olanzapine in childhood/adolescent pervasive developmental disorder. J Child Adolesc Psychopharmacol. 2006;16:541-548.
123. Research Units on Pediatric Psychopharmacology Autism Network. Risperidone in children with autism and serious behavioral problems. N Engl J Med. 2002;347:314-321.
124. Shea S, Turgay A, Carroll A, *et al.* Risperidone in the treatment of disruptive behavioral symptoms in children with autistic and other pervasive developmental disorders. Pediatrics. 2004;114: e634-e641.
125. McDougle CJ, Scahill L, Aman MG, *et al.* Risperidone for the core symptom domains of autism: results from the study by the Autism Network of the Research Units on Pediatric Psychopharmacology. Am J Psychiatry. 2005;162:1142-1148.
126. Miral S, Gencer O, Inal-Emiroglu FN, *et al.* Risperidone versus haloperidol in children and adolescents with AD: a randomized, controlled, double-blind trial. Eur Child Adolesc Psychiatry. 2008;17:1-8.
127. Hellings JA, Weekbaugh M, Nickel EJ, *et al.* A double-blinded placebo-controlled study of valproate for aggression in youth with pervasive developmental disorders. J Child Adolesc Psychopharmacol. 2005;15:682-692.
128. Hollander E, Soorya L, Wasserman S, *et al.* Divalproex sodium vs. placebo in the treatment of repetitive behaviours in autism spectrum disorder. Int J Neuropsychopharmacol. 2005;9:209-213.
129. Hollander E, Chaplin W, Soorya L, *et al.* Divalproex sodium vs. placebo for the treatment of irritability in children and adolescents with autism spectrum disorders. Neuropsychopharmacology. 2010;35:990-998.
130. Belsito L, Law P, Kirk K, *et al.* Lamotrigine therapy for autistic disorder: a randomized double-blind placebo-controlled trial. J Autism Dev Disord. 2001;31:175-181.
131. Wasserman S, Iyengar R, Chaplin WF, *et al.* Levetiracetam versus placebo in childhood and adolescent autism: a double-blind placebo-controlled study. Int Clin Psychopharmacol. 2006;21: 363-367.
132. Harfterkamp M, van de Loo-Neus G, Minderaa RB, *et al.* A randomized double-blind study of atomoxetine versus placebo for attention-deficit/hyperactivity disorder symptoms in children with autism spectrum disorder. J Am Acad Child Adolesc Psychiatry. 2012;51:733-741.
133. Arnold LE, Aman MG, Cook AM, *et al.* Atomoxetine for hyperactivity in autism spectrum disorders: placebo-controlled crossover pilot trial. J Am Acad Child Adolesc Psychiatry. 2006;45: 1196-1205.
134. King BH, Hollander E, Sikich L, *et al.* for the STAART Psychopharmacology Network. Lack of efficacy of citalopram in children with autism spectrum disorders and high levels of repetitive behavior: citalopram ineffective in children with autism. Arch Gen Psychiatry. 2009;66:583-590.
135. Hollander E, Phillips A, Chaplin W, *et al.* A placebo controlled crossover trial of liquid fluoxetine on repetitive behaviors in childhood and adolescent autism. Neuropsychopharmacology. 2005;30:582-589.
136. Gordon CT, State RC, Nelson JE. A double-blind comparison of clomipramine, desipramine, and placebo in the treatment of autistic disorder. Arch Gen Psychiatry. 1993;50:441-447.
137. Remington G, Sloman L, Konstantareas M, *et al.* Clomipramine versus haloperidol in the treatment of autistic disorder: a double-blind, placebo-controlled, crossover study. J Clin Psychopharmacol. 2001;4:440-444.
138. Research Units on Pediatric Psychopharmacology (RUPP) Autism Network. Randomized, controlled, crossover trial of methylphenidate in pervasive developmental disorders with hyperactivity. Arch Gen Psychiatry. 2005;62:1266-1274.
139. Pearson D, Santos CW, Aman MG, *et al.* Effects of extended release methylphenidate treatment on ratings of ADHD and associated behavior in children with autism spectrum disorders and ADHD symptoms. J Child Adolesc Psychopharmacology. 2013;23:337-351.
140. Handen BL, Johnson CR, Lubetsky M. Efficacy of methylphenidate among children with autism and symptoms of attention-deficit hyperactivity disorder. J Autism Dev Disord. 2000;30: 245-255.
141. Quintana H, Birmaher B, Stedge D, *et al.* Use of methylphenidate in the treatment of children with autistic disorder. J Autism Dev Disord. 1995;25:283-294.
142. King BH, Wright DM, Handen BL, *et al.* Double-blind, placebo-controlled study of amantadine hydrochloride in the treatment of children with autistic disorder. J Am Acad Child Adolesc Psychiatry. 2001;40:658-665.
143. Akhonzadeh S, Erfani S, Mohammadi M-R, *et al.* Cyproheptadine in the treatment of autistic disorder: a double-blind placebo-controlled trial. J Clin Pharm Ther. 2004;29:145-150.
144. Chez MG, Buchanan TM, Becker M, *et al.* Donepezil hydrochloride: a double-blind study in autistic children. J Pediatr Neurol. 2003;1:83-88.
145. Willemsen-Swinkels SH, Buitelaar JK, Nijhof GJ, *et al.* Failure of naltrexone hydrochloride to reduce self-injurious and autistic behavior in mentally retarded adults: double-blind placebo-controlled studies. Arch Gen Psychiatry. 1995;52:766-773.
146. Kolmen BK, Feldman HM, Handen BL, *et al.* Naltrexone in young autistic children: a double-blind, placebo-controlled crossover study. J Am Acad Child Adolesc Psychiatry. 1995;34 (2):223-231.
147. Feldman HM, Kolmen BK, Gonzaga AM. Naltrexone and communication skills in young children with autism. J Am Acad Child Adolesc Psychiatry. 1999;38:587-593.
148. Campbell M, Anderson LT, Small AM, *et al.* Naltrexone in autistic children: behavioral symptoms and attentional learning. J Am Acad Child Adolesc Psychiatry. 1993;32:1283-1291.
149. Campbell M, Anderson LT, Small AM, *et al.* Naltrexone in autistic children: a double-blind and placebo-controlled study. Psychopharmacol Bull. 1990;26:130-135.
150. Akhonzadeh S, Fallah J, Mohammadi M-R, *et al.* Double-blind placebo-controlled trial of pentoxifylline added to risperidone: effects on aberrant behavior in children with autism. Prog Neuropsychopharmacol Biol Psychiatry. 2010;34:32-36.
151. Scahill L, McDougle CJ, Aman MG, *et al.* Effects of risperidone and parent training on adaptive functioning in children with pervasive developmental disorders and serious behavioral problems. J Am Acad Child Adolesc Psychiatry. 2012;51:136-146.
152. Aman MG, Mcdougle CJ, Scahill L, *et al.* Medication and parent training in children with pervasive developmental disorders and serious behavioral problems: results from a randomized clinical trial. J Am Acad Child Adolesc Psychiatry. 2009;48: 1143-1154.
153. Aman MG, Novotny S, Samango-Sprouse C, *et al.* Outcome measures for clinical drug trials in autism. CNS Spectr. 2004; 9:36-47.
154. Freedman BH, Kalb LG, *et al.* Relationship status among parents of children with autism spectrum disorders: a population-based study. J Autism Dev Disord. 2012;42:539-548.
155. Wong HHL, Smith RG. Patterns of complementary and alternative medical therapy use in children diagnosed with autism spectrum disorders. J Autism Dev Disord. 2006;36:901-909.
156. Williams KJ, Wray JJ, Wheeler DM. Intravenous secretin for autism spectrum disorders. Cochrane Database Syst Rev. 2005;3: CD003495.
157. Findling RL, Scotese-Wojtila L, Huang J, *et al.* High-dose pyridoxine and magnesium administration in children with autistic disorder: An absence of salutary effects in a double-blind, placebo-controlled study. J Autism Dev Disord. 1997;27:467-478.
158. Milward C, Ferriter M, Calver S, *et al.* Gluten- and casein-free diets for autistic spectrum disorder. Cochrane Database Syst Rev. 2008;2;CD003498.

159. James S, Montgomery P, Williams K. Omega-3 fatty acids supplementation for autism spectrum disorders (ASD). Cochrane Database Syst Rev. 2011;11:CD007992.
160. Handen BL, Melmed RD, Hansen RL, *et al.* A double-blind, placebo-controlled trial of oral human immunoglobulin for gastrointestinal dysfunction in children with autistic disorder. J Autism Dev Disord. 2009;39:796-805.
161. Brown MJ, Willis T, Omalu B, Leiker R. Deaths resulting from hypocalcemia after administration of edetate disodium: 2003-2005. Pediatrics. 2006;118:e534-e536.
162. Jacobson JW, Foxx RM, Mulick JA. Controversial Therapies for Developmental Disabilities: Fad, Fashion and Science in Professional Practice. Mahwah, NJ: Lawrence Erlbaum Associates; 2005.
163. Levy S, Hyman S. Dietary, complementary, and alternative therapies. In: Riechow B, Doehring P, Cichetti D, Volkmar F, eds. Evidence Based Practices and Treatments for Children with Autism. New York: Springer; 2011:275-286.
164. Reichow B, Peohring P, Cocchetti DM, Volkmar FR, eds. Evidence Based Practices and Treatments for Children with Autism. New York: Springer; 2011.

Journal of Clinical Child and Adolescent Psychology
2005, Vol. 34, No. 3, 506–522

Evidence-Based Assessment of Learning Disabilities in Children and Adolescents

Jack M. Fletcher
Department of Pediatrics and the Center for Academic and Reading Skills, University of Texas Health Science Center at Houston

David J. Francis
Department of Psychology and the Texas Institute for Measurement, Evaluation and Statistics, University of Houston

Robin D. Morris
Department of Psychology, Georgia State University

G. Reid Lyon
Child Development and Behavior Branch, National Institute of Child Health and Human Development

The reliability and validity of 4 approaches to the assessment of children and adolescents with learning disabilities (LD) are reviewed, including models based on (a) aptitude–achievement discrepancies, (b) low achievement, (c) intra-individual differences, and (d) response to intervention (RTI). We identify serious psychometric problems that affect the reliability of models based on aptitude–achievement discrepancies and low achievement. There are also significant validity problems for models based on aptitude–achievement discrepancies and intra-individual differences. Models that incorporate RTI have considerable potential for addressing both the reliability and validity issues but cannot represent the sole criterion for LD identification. We suggest that models incorporating both low achievement and RTI concepts have the strongest evidence base and the most direct relation to treatment. The assessment of children for LD must reflect a stronger underlying classification that takes into account relations with other childhood disorders as well as the reliability and validity of the underlying classification and resultant assessment and identification system. The implications of this type of model for clinical assessments of children for whom LD is a concern are discussed.

Assessment methods for identifying children and adolescents with learning disabilities (LD) are multiple, varied, and the subject of heated debates among practitioners. Those debates involve issues that extend beyond the value of specific tests, often reflecting different views of how LD is best identified. These views reflect variations in the definition of LD and, therefore, variations in what measures are selected to operationalize the definition (Fletcher, Foorman, et al., 2002). Any focus on the "best tests" leads to a hopeless morass of confusion in an area such as LD that has not successfully addressed the classification and definition issues that lead to identification of who does and who does not possesses characteristics of LD. Definitions always reflect an implicit classification indicating how different constructs are measured and used to identify members of the class in terms of similarities and differences relative to other entities that are not considered members of the class (Morris & Fletcher, 1988). For LD, children who are members of this class are historically differentiated from children who have other achievement-related difficulties, such as mental retardation, sensory disorders, emotional or behavioral disturbances, and environmental causes of underachievement, including economic disadvantage, minority language status, and inadequate instruction (Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1993; Lyon, Fletcher, & Barnes, 2003). If the classification is valid, children with LD may share characteristics that are similar with other groups of underachievers, but they

Grants from the National Institute of Child Health and Human Development, P50 21888, Center for Learning and Attention Disorders, and National Science Foundation 9979968, Early Reading Development: A Cognitive Neuroscience Approach supported this article.

We gratefully acknowledge contributions of Rita Taylor to preparation of this article.

Requests for reprints should be sent to Jack M. Fletcher, Department of Pediatrics, University of Texas Health Science Center at Houston, 7000 Fannin Street, UCT 2478, Houston, TX 77030. E-mail: Jack.Fletcher@uth.tmc.edu

should also differ in ways that can be measured and that can serve to define and operationalize the class of children and adolescents with LD.

In this article, we consider evidence-based approaches to the assessment of LD in the context of different approaches to the classification and identification of LD. We argue that the measurement systems that are used to identify children and adolescents with LD are inseparable from the classifications from which the identification criteria evolve. Moreover, all measurement systems are imperfect attempts to measure a construct (LD) that operates as a latent variable that is unknowable independently of how it is measured and therefore of how LD is classified. The construct of LD is imperfectly measured simply because the measurement tools themselves are not error free (Francis et al., 2005). Different approaches to classification and definition capitalize on this error of measurement in ways that reduce or increase the reliability of the classification itself. Similarly, evaluating similarities and differences among groups of students who are identified as LD and not LD is a test of the validity of the underlying classification, so long as the variables used to assess this form of validity are not the same as those used for identification (Morris & Fletcher, 1988). As with any form of validity, adequate reliability is essential. Classifications can be reliable and still lack validity. The converse is not true; they cannot be valid and lack reliability. A valid classification of LD predicts important characteristics of the group. Consistent with the spirit of this special section, the most important characteristic is whether the classification is meaningfully related to intervention. For LD, a classification should also predict a variety of differences on cognitive skills, behavioral attributes, and achievement variables not used to form the classification, developmental course, response to intervention (RTI), neurobiological variables, or prognosis (Fletcher, Lyon, et al., 2002).

To address these issues, we consider the reliability and validity of four approaches to the classification and assessment of LD: (a) IQ discrepancy and other forms of aptitude–achievement discrepancy, (b) low achievement, (c) intra-individual differences, and (d) models incorporating RTI and some form of curriculum-based measurement. We consider how each classification reflects the historically prominent concept of "unexpected underachievement" as the key construct in LD assessment (Lyon et al., 2001), that is, what many early observers characterized as a group of children unable to master academic skills despite the absence of known causes of poor achievement (sensory disorder, mental retardation, emotional disturbances, economic disadvantages, inadequate instruction). From this perspective, a valid classification and measurement system for LD must identify a unique group of underachievers that is clearly differentiated from groups with other forms of underachievement.

Defining LD

Historically, definition and classification issues have haunted the field of LD. As reviewed in Lyon et al. (2001), most early conceptualizations viewed LD simply as a form of "unexpected" underachievement. The primary approach to assessment involved the identification of intra-individual variability as a marker for the unexpectedness of LD, along with the exclusion of other causes of underachievement that would be expected to produce underachievement. This type of definition was explicitly coded into U.S. federal statutes when LD was identified as an eligibility category for special education in Public Law 94–142 in 1975; essentially the same definition is part of current U.S. federal statues in the Individuals with Disabilities Education Act (1997).

The U.S. statutory definition of LD is essentially a set of concepts that in itself is difficult to operationalize. In 1977, recommendations for operationalizing the federal definition of LD were provided to states after passage of Public Law 94–142 to help identify children in this category of special education (U. S. Office of Education, 1977). In these regulations, LD was defined as a heterogeneous group of seven disorders (oral language, listening comprehension, basic reading, reading comprehension, math calculations, math reasoning, written language) with a common marker of intra-individual variability represented by a discrepancy between IQ and achievement (i.e., unexpected underachievement). Unexpectedness was also indicated by maintaining the exclusionary criteria present in the statutory definition that presumably lead to expected underachievement. Other parts of the regulations emphasize the need to ensure that the child's educational program provided adequate opportunity to learn. No recommendations were made concerning the assessment of psychological processes, most likely because it was not clear that reliable methods existed for assessing processing skills and because the field was not clear on what processes should be assessed (Reschly, Hosp, & Smied, 2003).

This approach to definition is now widely implemented with substantial variability across schools, districts, and states in which students are served in special education as LD (MacMillan & Siperstein, 2002; Mercer, Jordan, Allsop, & Mercer, 1996; Reschly et al., 2003). It is also the basis for assessments of LD outside of schools. Consider, for example, the definition of reading disorders in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994), which indicates that the student must perform below levels expected for age and IQ, and specifies only sensory disorders as exclusionary:

A. Reading achievement, as measured by individually administered standardized tests of read-

ing accuracy or comprehension, is substantially below that expected given the person's chronological age, measured intelligence, and age-appropriate education.

B. The disturbance in Criterion A significantly interferes with academic achievement or activities of daily living that require reading skills.

C. If a sensory deficit is present, the reading difficulties are in excess of those usually associated with it.

The International Classification of Diseases–10 has a similar definition. It differs largely in being more specific in requiring use of a regression-adjusted discrepancy, specifying cut points (achievement two standard errors below IQ) for identifying a child with LD, and expanding the range of exclusions.

Although these definitions are used in what are often disparate realms of practice, they lead to similar approaches to the identification of children and adolescents as LD. Across these realms, children commonly receive IQ and achievement tests. The IQ test is commonly interpreted as an aptitude measure or index against which achievement is compared. Different achievement tests are used because LD may affect achievement in reading, math, or written language. The heterogeneity is recognized explicitly in the U.S. statutory and regulatory definitions of LD (Individuals With Disabilities Education Act, 1997) and in the psychiatric classifications by the provision of separate definitions for each academic domain. However, it is still essentially the same definition applied in different domains. In many settings, this basic assessment is supplemented with tests of processing skills derived from multiple perspectives (neuropsychology, information processing, and theories of LD). The approach boils down to administration of a battery of tests to identify LD, presumably with treatment implications.

Underlying Classification Hypotheses

Implicit in all these definitions are slight variations on a classification model of individuals with LD as those who show a measurable discrepancy in some but not all domains of skill development and who are not identified into another subgroup of poor achievers. In some instances, the discrepancy is quantified with two tests in an aptitude–achievement model epitomized by the IQ-discrepancy approach in the U.S. federal regulatory definition and the psychiatric classifications of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) and the International Classification of Diseases–10. Here the classification model implicitly stipulates that those who meet an IQ-discrepancy inclusionary criterion are different in meaningful ways from those who are underachievers and do not meet the discrepancy criteria or criteria for one of the exclusionary conditions. Some have argued that this model lacks validity and propose that LD is synonymous with underachievement, so that it should be identified solely by achievement tests (Siegel, 1992), often with some exclusionary criteria to help ensure that the achievement problem is unexpected. Thus, the contrast is really between a two-test aptitude–achievement discrepancy and a one-test chronological age-achievement discrepancy with achievement low relative to age-based (or grade-based) expectations. If processing measures are added, the model becomes a multitest discrepancy model. Identification of a child as LD in all three of these models is typically based on assessment at a single point in time, so we refer to them as "status" models. Finally, RTI models emphasize the "adequate opportunity to learn" exclusionary criterion by assessing the child's response to different instructional efforts over time with frequent brief assessments, that is, a "change" model. The child who is LD becomes one who demonstrates intractability in learning characteristics by not responding adequately to instruction that is effective with most other students.

Dimensional Nature of LD

Each of these four models can be evaluated for reliability and validity. Unexpected underachievement, a concept critically important to the validity of the underlying construct of LD, can also be examined. The reliability issues are similar across the first three models and stem from the dimensional nature of LD. Most population-based studies have shown that reading and math skills are normally distributed (Jorm, Share, Matthews, & Matthews, 1986; Lewis, Hitch, & Walker, 1994; Rodgers, 1983; Shalev, Auerbach, Manor, & Gross-Tsur, 2000; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992; Silva, McGee, & Williams, 1985). These findings are buttressed by behavioral genetic studies, which are not consistent with the presence of qualitatively different characteristics associated with the heritability of reading and math disorders (Fisher & DeFries, 2002; Gilger, 2002). As dimensional traits that exist on a continuum, there would be no expectation of natural cut points that differentiate individuals with LD from those who are underachievers but not identified as LD (Shaywitz et al., 1992).

The unobservable nature of LD makes two-test and one-test discrepancy models unreliable in ways that are psychometrically predictable but not in ways that simply equate LD with poor achievement (Francis et al., 2005; Stuebing et al., 2002). The problem is that the measurement approach is based on a static assessment model that possesses insufficient information about the underlying construct to allow for reliable classifications of individuals along what is essentially an unobservable dimension. If LD was a

manifest concept that was directly observable in the behavior of affected individuals, or if there were natural discontinuities that represented a qualitative breakpoint in the distribution of achievement skills or the cognitive skills on which achievement depends, this problem would be less of an obstacle. However, like achievement or intelligence, LD is a latent construct that must be inferred from the pattern of performance on directly observable operationalizations of other latent constructs (namely, test scores that index constructs like reading achievement, phonological awareness, aptitude, and so on). The more information available to support the inference of LD, the more reliable (and valid) that inference becomes, thus supporting the fine-grained distinctions necessitated by two-test and one-test discrepancy models. To the extent that the latent construct, LD, is categorical, by which we mean that the construct indexes different classes of learners (i.e., children who learn differently) as opposed to simply different levels of achievement, then systems of identification that rely on one measurable variable lack sufficient information to identify the latent classes and assign individuals to those classes without placing additional, untestable, and unsupportable constraints on the system. It is simply not possible to use a single mean and standard deviation and to estimate separate means and standard deviations for two (or more) unobservable latent classes of individuals and determine the percentage of individuals falling into each class, let alone to classify specific individuals into those classes. Without constraints, such as specifying the magnitude of differences in the means of the latent classes, the ratio of standard deviations, and the odds of membership in the two (or more) classes, the system is under-identified, which simply means that there are many different solutions that cannot be distinguished from one another.

When the system is under-identified, the only solution is to expand the measurement system to increase the number of observed relations, which in one sense is what intra-individual difference models attempt by adding assessments of processing skills. Other criteria are necessary because it is impossible to uniquely identify a distinct subgroup of underachieving individuals consistent with the construct of LD when identification is based on a single assessment at a single time point. Adding external criteria, such as an aptitude measure or multiple assessments of processing skills, increases the dimensionality of the measurement system and makes latent classification more feasible, even when the other criteria are themselves imperfect. But the main issues for one-test, two-test, and multitest identification models involve the reliability of the underlying classifications and whether they identify a unique subgroup of underachievers. In the next section, we examine variations in reliability and validity for each of these models, focusing on the importance of reliability, as the validity of the classifications can be no stronger than their reliability.

Models Based on Two-Test Discrepancies

Although the IQ-discrepancy model is the most widely utilized approach to identifying LD, there are many different ways to operationalize the model. For example, some implementations are based on a composite IQ score, whereas others utilize either a verbal or nonverbal IQ score. Other approaches drop IQ as the aptitude measure and use a measure such as listening comprehension. In the validity section, we discuss each of these approaches. The reliability issues are similar for each example of an aptitude–achievement discrepancy.

Reliability

Specific reliability problems for two-test discrepancy models pertain to any comparison of two correlated assessments that involve the determination of a child's performance relative to a cut point on a continuous distribution. Discrepancy involves the calculation of a difference score (D) to estimate the true difference (Δ) between two latent constructs. Thus, discussions about discrepancy must distinguish between problems with the manifest (i.e., observed) difference (D) as an index of the true difference (Δ) but also must consider whether the true difference (Δ) reflects the construct of interest. Problems with the reliability of D based on differences between two tests are well known, albeit not in the LD context (Bereiter, 1967). However, there is nothing that fundamentally limits the applicability of this research to LD if we are willing to accept a notion of Δ as a marker for LD. There are major problems with this assumption that are reviewed in Francis et al. (2005). The most significant is regression to the mean. On average, regression to the mean indicates that scores that are above the mean will be lower when the test is repeated or when a second correlated test is used to compute D. In this example, individuals who have IQ scores above the mean will obtain achievement test scores that, on average, will be lower than the IQ test score because the achievement score will move toward the mean. The opposite is true for individuals with IQ scores below the mean. This leads to the paradox of children with achievement scores that exceed IQ, or the identification of low-achieving, higher IQ children with achievement above the average range as LD.

Although adjusting for the correlation of IQ and achievement helps correct for regression effects (Reynolds, 1984–1985), unreliability also stems from the attempt to assess a person's standing relative to a cut point on a continuous distribution. As discussed in the

following section on low achievement models, this problem makes identification with a single test—even one with small amounts of measurement error—potentially unreliable, a problem for any status model.

None of this discussion addresses the validity question concerning Δ. Specifically, does Δ embody LD as we would want to conceptualize it (e.g., as unexpected underachievement), or is Δ merely a convenient conceptualization of LD because it is a conceptualization that leads directly to easily implemented, operational definitions, however flawed they might be?

Validity

The validity of the IQ-discrepancy model has been extensively studied. Two independent meta-analyses have shown that effect sizes on measures of achievement and cognitive functions are in the negligible to small range (at best) for the comparison of groups formed on the basis of discrepancies between IQ and reading achievement versus poor readers without an IQ discrepancy (Hoskyn & Swanson, 2000; Stuebing et al., 2002), findings similar to studies not included in these meta-analyses (Stanovich & Siegel, 1994). Other validity studies have not found that discrepant and nondiscrepant poor readers differ in long-term prognosis (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Silva et al., 1985), response to instruction (Fletcher, Lyon, et al., 2002; Jiménez et al., 2003; Stage, Abbott, Jenkins, & Berninger, 2003; Vellutino, Scanlon, & Jaccard, 2003), or neuroimaging correlates (Lyon et al., 2003; but also see Shaywitz et al., 2003, which shows differences in groups varying in IQ but not IQ discrepancy). Studies of genetic variability show negligible to small differences related to IQ-discrepancy models that may reflect regression to the mean (Pennington, Gilger, Olson, & DeFries, 1992; Wadsworth, Olson, Pennington, & DeFries, 2000). Similar empirical evidence has been reported for LD in math and language (Fletcher, Lyon, et al., 2002; Mazzocco & Myers, 2003). This is not surprising given that the problems are inherent in the underlying psychometric model and have little to do with the specific measures involved in the model except to the extent that specific test reliabilities and intertest correlations enter into the equations.

Despite the evidence of weak validity for the practice of differentiating discrepant and nondiscrepant students, alternatives based on discrepancy models continue to be proposed, and psychologists outside of schools commonly implement this flawed model. However, given the reliability problems inherent in IQ discrepancy models, it is not surprising that these other attempts to operationalize aptitude–achievement discrepancy have not met with success. In the Stuebing et al. (2002) meta-analysis, 32 of the 46 major studies had a clearly defined aptitude measure. Of these studies, 19 used Full Scale IQ, 8 used Verbal IQ, 4 used Performance IQ, and 1 study used a discrepancy of listening comprehension and reading comprehension. Not surprisingly, these different discrepancy models did not yield results that were different from those when a composite IQ measure was utilized. Neither Fletcher et al. (1994) nor Aaron, Kuchta, and Grapenthin (1988) were able to demonstrate major differences between discrepant and low achievement groups formed on the basis of listening comprehension and reading comprehension.

The differences in these models involve slight changes in who is identified as discrepant or low achieving depending on the cut point and the correlation of the aptitude and achievement measures. The changes simply reflect fluctuations around the cut point where children are most similar. It is not surprising that effect sizes comparing poor achievers with and without IQ discrepancies are uniformly low across these different models. Current practices based on this approach to identification of LD epitomized by the federal regulatory definition and psychiatric classifications are fundamentally flawed.

One-Test (Low Achievement) Models

Reliability

The measurement problems that emerge when a specific cut point is used for identification purposes affect *any* psychometric approach to LD identification. These problems are more significant when the test score is not criterion referenced, or when the score distributions have been smoothed to create a normal univariate distribution. To reiterate, the presence of a natural breakpoint in the score distribution, typically observed in multimodal distributions, would make it simple to validate cut points. But natural breaks are not usually apparent in achievement distributions because reading and math achievement distributions are normal. Thus, LD is essentially a dimensional trait, or a variation on normal development.

Regardless of normality, measurement error attends any psychometric procedure and affects cut points in a normal distribution (Shepard, 1980). Because of measurement error, any cut point set on the observed distribution will lead to instability in the identification of class members because observed test scores will fluctuate around the cut point with repeated testing or use of an alternative measure of the same construct (e.g., two reading tests). This fluctuation is not just a problem of correlated tests or simply a matter of setting better cut scores or developing better tests. Rather, no single observed test score can capture perfectly a student's ability on an imperfectly measured latent variable. The fluctuation in identifications will vary across different tests, depending in part on the measurement

error. In both real and simulated data sets, fluctuations in up to 35% of cases are found when a single test is used to identify a cut point. Similar problems are apparent if a two-test discrepancy model is used (Francis et al., 2005; Shaywitz et al., 1992).

This problem is less of an issue for research, which rarely hinges on the identification of individual children. Thus, it does not have great impact on the validity of a low achievement classification because, on average, children around the cut point who may be fluctuating in and out of the class of interest with repeated testing are not very different. However, the problems for an individual child who is being considered for special education placement or a psychiatric diagnosis are obvious. A positive identification in either example often carries a poor prognosis.

Validity

Models based on the use of achievement markers can be shown to have a great deal of validity (see Fletcher, Lyon, et al., 2002; Fletcher, Morris, & Lyon, 2003; Siegel, 1992). In this respect, if groups are formed such that the participants do not meet criteria for mental retardation and have achievement scores that are below the 25th percentile, a variety of comparisons show that subgroups of underachievers emerge that can be validly differentiated on external variables and help demonstrate the viability of the construct of LD. For example, if children with reading and math disabilities identified in this manner are compared to typical achievers, it is possible to show that these three groups display different cognitive correlates. In addition, neurobiological studies show that these groups differ both in the neural correlates of reading and math performance as well as the heritability of reading and math disorders (Lyon et al., 2003). These achievement subgroups, which by definition include children who meet either low achievement or IQ-discrepancy criteria, even differ in RTI, providing strong evidence for "aptitude by treatment" interactions; math interventions provided for children with reading problems are demonstrably ineffective, and vice versa.

Despite this evidence for validity, concerns emerge about definitions based solely on achievement cut points. Simply utilizing a low achievement definition, even when different exclusionary criteria are applied, does not operationalize the true meaning of unexpected underachievement. Although such an approach to identification is deceptively simple, it is arguable whether the subgroups that remain represent a unique group of underachievers. For example, how well are underachievers whose low performance is attributed to LD differentiated from underachievers whose low performance is attributed to emotional disturbance, economic disadvantage, or inadequate instruction (Lyon et al., 2001)? To use the example of word recognition, there is little evidence that these subgroups vary in terms of phonological awareness or other language tasks, RTI, or even neuroimaging correlates. In this respect, the validity is weak because the underlying construct of LD is not adequately assessed. Additional criteria are needed, but simply adding a single aptitude measure *decreases* reliability and does not add to the validity of a low achievement definition.

Models Based on Intra-Individual Differences

A commonly proposed alternative to models based on aptitude–achievement discrepancies or low achievement involves an examination of individual differences on measures of cognitive function. Thus, for example, a recent consensus article from 10 major advocacy groups organized by the National Center for Learning Disabilities (2002) stated that "while IQ tests do not measure or predict a student's response to instruction, measures of neuropsychological functioning and information processing could be included in evaluation protocols in ways that document the areas of strength and vulnerability needed to make informed decisions about eligibility for services, or more importantly, what services are needed. An essential characteristic of LD is failure to achieve at a level of expected performance based upon the student's other abilities" (p. 4).

This statement proposes intra-individual differences as a marker for unexpected underachievement. As opposed to a single marker such as IQ discrepancy or low achievement, unexpectedness is operationalized as unevenness in scores across multiple tests. The person identified as LD (by definition) has strengths in many areas of cognitive or neuropsychological function but weaknesses in core attributes that lead to underachievement. The LD is unexpected because the weaknesses lead to selected and narrow difficulties with achievement and adaptive functions. Proponents of this view believe that such approaches identify children as LD based on profiles across tests that differentiate types of LD and also differentiate LD from other childhood disorders, such as mental retardation and behavioral disorders such as attention deficit hyperactivity disorder (ADHD). This approach leads to definitions based on inclusionary criteria in which children are identified as LD based on characteristics that relate to intra-individual differences (Lyon et al., 2001).

Reliability

In essence, the intra-individual difference model employs a multitest discrepancy approach and carries with it the problems involved with estimation of discrepancies and cut points. These problems are inherent in any attempt to identify a person as LD (Fletcher et

al., 2003). However, examining patterns of test scores has long been favored by clinical neuropsychologists, largely because it seems to correspond more closely with clinical practice and because it adds information to the decision-making process (see the elegant discussion of differential test scores, discrepancies, and profiles in Rourke, 1975). The unique reliability issue involves the idea that LD is represented by unevenness in test profiles. This may be true, but does this observation mean that children with flatter profiles are not LD? Severity is correlated with the shape of a profile due to the lack of independence of different tests that might be used to construct the profile (Morris, Fletcher, & Francis, 1993). Children with increasingly severe reading problems, for example, will show increasingly flat profiles across processing measures (e.g., phonological awareness, rapid naming, and vocabulary) in direct correspondence to severity because all these measures are moderately correlated. Thus, if the inclusionary criterion for the presence of LD is evidence of a discrepancy in neuropsychological or processing skills, such an approach may exclude the most severely impaired children, irrespective of global measures such as IQ, because more severely impaired children are less likely to show skill discrepancies due to the intercorrelation of the tests (Morris et al., 1993, 1998).

Validity

A major assumption of a multitest intra-individual differences model is that identification based on performance patterns will lead to enhanced treatment of children with LD. It is commonly assumed that such tests point out areas that need intervention. However, there is little evidence that strengths and weaknesses in processing skills are related to intervention outcomes. It is well established that training in underlying processes does not usually generalize into the related academic area (Lyon & Moats, 1988; Reschly, Tilly, & Grimes, 1999; Vellutino, 1979). For example, training on phonological awareness skills without explicit transfer to a letter component produces gains in phonological awareness but not in reading (National Reading Panel, 2000). Training in auditory or visual perceptual skills does not lead to better outcomes for children identified as "auditory" or "visual" learners (Lyon, Fletcher, Fuchs, & Chhabra, in press; Vellutino, Fletcher, Scanlon, & Snowling, 2004).

There is support for the idea that intra-individual differences identify some children as LD, epitomized by the link of dyslexia with word recognition and phonological processing (Vellutino et al., 2004). Even here the intra-individual differences model focuses on skills that are only correlated with the achievement domain. Simply identifying children with LD based solely on processing skills is questionable and would likely yield many false positive identifications of children as LD without achievement difficulties (Torgesen, 2002). The reliability of many processing measures is lower than those associated with the achievement (or IQ) domain, so such false positives should be expected. Other than the word recognition–phonological processing link, relations of processing and other forms of LD are not well established (Torgesen, 2002). Finally, what do we learn about variability in processing skills that is not apparent in profiles across achievement domains (Fletcher et al., 2003)? In fact, the model has the most validity at the level of achievement markers but simply collapses into a low achievement model in the absence of processing measures. Thus, if we accept the notion that specific discrepancies in cognitive domains are a unique marker for LD, given that the processing measures are usually linked to an achievement domain, what is unique about variations in processing skills that is not apparent in variations in achievement domains? Would we eliminate as LD students who have difficulties in reading, math, and writing? This is not viable, as impairments in all domains often occur in non-mentally retarded children with language-based difficulties.

Models Incorporating RTI

An alternative approach to status models that would increase the reliability of these would increase the number of time points whereby a child was assessed. Shepard (1980), for example, proposed that IQ discrepancies could be assessed more reliably if a child was tested four times. The impracticality of such an approach, which would require about 10 to 12 hr per child, is obvious, not to mention that even more resources would be devoted to determination of eligibility, taking away funds and time needed for intervention.

Another approach to increasing the number of time points would involve much shorter assessments of key achievement skills over time. These approaches, or RTI models, typically involve identification practices based in part on multiple short assessment probes of knowledge and performance in a specific academic domain, such as reading or math (Fuchs & Fuchs, 1998). By linking multiple assessments to specific attempts to intervene with the child, the construct of unexpected underachievement can be operationalized, in part, on the basis of nonresponsiveness to instruction to which most other students respond (Gresham, 2002). In fact, this is still a variation of a discrepancy model, but the advantage is that the model is better identified because of multiple short assessments of a key attribute (e.g., reading, math) over time.

Such models have been proposed in several recent consensus reports that address LD identification (Bradley, Danielson, & Hallahan, 2002; President's

Commission on Excellence in Special Education, 2002), most notably in a recent report of the National Research Council (Donavon & Cross, 2002). These reports suggest that one criterion for LD identification is when a student does not respond to high-quality instruction and intervention. The implementation of this approach requires frequent monitoring of progress as the student receives the intervention (Fuchs & Fuchs, 1998). Speece and Case (2001) found that serial assessments of growth and level of performance in reading fluency predicted reading problems in at-risk children better than a single assessment of fluency. This approach is anchored in a system known as curriculum-based measurement, where the assessments themselves have adequate reliability and are constantly improving (Fuchs & Fuchs, 1999; Shinn, 1998).

Reliability

Are RTI approaches that involve multiple assessments over time psychometrically more reliable than traditional approaches to LD identification? An approach based on multiple measures over time has the potential to reduce the difficulties encountered with reliance on a single assessment at a single time point. Certainly the reliability of the multiple assessment approach is greater than if the single assessment is used to form a discrepancy, because typically the discrepancy will be a poorer (i.e., less reliable) measure of the true difference than the observed measures of their respective underlying constructs. Focusing on successive measurements over time has the effect of moving the identification process from "ability–ability" comparisons (two different abilities compared at one point in time) to "ability change" models (same ability over time). Such approaches have the potential to ameliorate the difficulties associated with ability–ability discrepancies, whether univariate or bivariate, because they involve the use of more than two assessment time points. Generally, the more information that is brought to bear on any eligibility or diagnostic decision, the more reliable the decision, although it is certainly possible to create counterexamples by combining information from irrelevant or confounding sources. Such irrelevancies are not likely to be introduced by assessing the same skill over time as in a model that incorporates RTI, when that skill was previously deemed relevant to assess at a single time point.

Conceptually, the study of change is made more feasible by the collection of multiple assessments because the precision by which change can be measured improves as the number of time points increases (Rogosa, 1995). When more than two assessment time points are collected, the reliability of estimated change can also be estimated directly from the data, and the imprecision inherent in individual estimates can be used to provide improved estimates of growth parameters for individual students as well as for groups of students. If change is not linear, the use of four or more time points can map the form of growth. And for those who favor status models over change or learning models, it remains possible to use the intercept term in the individual growth model as an estimate of status. This intercept provides a more precise estimate of true status at any single point in time than would any single assessment.

These approaches are not without difficulty. The introduction of serial assessments has not eliminated the necessity of indirect estimation of the parameters of interest. In the discrepancy model, D is used to estimate Δ. A model incorporating RTI uses a complex function of the observed data for individual i as well as the data from many other individuals to estimate each of the π_{ij}, the j true learning parameters for individual i. Different approaches to this estimation problem have varying strengths and weaknesses but all will make assumptions about the arithmetic form of the model, the distribution of the learning parameters, and the distributions of the errors. The ramifications of these assumptions for inferences about individual learning parameters must be studied in the LD context.

Models based on RTI also involve imperfect measures that include measurement error (Fletcher et al., 2003). However, this problem is reduced because of the use of multiple assessments and the borrowing of precision from the entire collection of data to provide a more precise estimate of the growth parameters of each individual. Thus, it becomes possible to estimate a child's "true" status more precisely as well as to estimate the rate of skill acquisition and to use these estimates as indicators of LD. In addition, this approach to estimation makes assumptions about the distribution of errors of measurement. In some cases, errors might be assumed to be uncorrelated. Again, this assumption must be examined in terms of its importance to inferences about individual status and rates of learning. In many cases, the inclusion of multiple assessment time points will allow this assumption to be relaxed, and the correlation among errors of measurement can be estimated and taken into account in forming inferences about individual status and rates of learning.

There still could be a need to identify individual children as LD based on cut points unless the entire process devolves to clinical judgment. Models that include RTI do not solve the issue of the dimensional versus categorical nature of LD. Determining cut points and benchmarks, for example, will continue to be an arbitrary process until cut points are linked to functional outcomes (Cisek, 2001), an issue never really addressed in LD identification for any identification model. However, models that include RTI have the promise of incorporating functional outcomes because they are tied to intervention response.

Validity

The introduction of serial assessments has an advantage beyond any statistical advantage it may confer for the estimation of individual's true status. Specifically, the introduction of serial assessments brings learning and the measurement of change to the forefront in conceptualizations of LD. The collection of serial assessments under specified conditions of effective instruction simultaneously focuses the definition of LD on a failure to learn, where learning can be measured more directly. Moreover, the specific instructional elements and the conditions under which they are implemented can be described, thereby providing a clearer basis for the expectation of learning and the unexpectedness of any failure to learn. Finally, focusing on multiple assessments in a RTI model has the advantage of clearly tying the identification process to the most important component of the construct of LD, which is unexpected underachievement. Models that incorporate RTI may identify a unique group of children that can be clearly differentiated from other low achievers in terms of cognitive correlates, prognosis, and even neurobiological factors.

Studies of children defined using different methods as responders and nonresponders clearly show significant differences in cognitive skills. For example, Stage et al. (2003), Vellutino et al. (2003), and Vaughn, Linan-Thompson, and Hickman-Davis (2003) found that nonresponders to early intervention differed from responders in both preintervention achievement scores and preintervention cognitive tasks. Nonresponders typically had more severe deficits in both reading-related factors (e.g., phonemic awareness, fluency) and reading skills. In recent imaging studies involving both early intervention and remediation of older students (see Fletcher, Simos, Papanicolaou, & Denton, 2004), we likewise found that individuals who were nonresponders showed more severe reading difficulties prior to intervention. More dramatic were the differences in neuroimaging correlates between those who responded to intervention and those who did not. We have found that nonresponders persist with a brain activation pattern that generally demonstrates a failure to activate left hemisphere areas known to be involved in the development of reading skills. In fact, those who were nonresponders showed predominant right-hemisphere activity much like that observed in children and adults with identified reading disabilities (Fletcher et al., 2004).

Implications for Clinical Assessments

This review of classification models may seem removed from the question of how to conduct clinical assessments of children suspected of LD. In fact, when a psychologist conducts any assessment for LD, the selection of tests reflects the underlying classification model and the constructs it specifies. If the psychologist or educator adopts an aptitude–achievement discrepancy model, the primary tools will be the tests used to operationalize aptitude (e.g., IQ) and achievement. If the clinician adopts a low achievement model, aptitude will not be measured—just achievement. An intra-individual differences model will require neuropsychological or cognitive processing measures. If a model is used that incorporates RTI, assessments of the integrity of the implementation of the intervention and progress monitoring assessments are necessary.

In evaluating models, we found little evidence that supports the aptitude–achievement and intra-individual difference models. Both involve the assessment of cognitive processes that do not contribute to the identification of a unique group of underachievers with LD and have serious reliability problems. The low achievement model has more reliability and validity but does not identify a unique group of underachievers. RTI criteria may permit identification of a unique group of underachievers but by themselves are not sufficient for identification of LD. Combining the strengths of the low achievement and RTI models leads to a hybrid model that invokes concepts of low achievement and RTI. This model can be expanded to incorporate assessment of contextual factors and other disorders that should be evaluated because of the need for differential treatment (Fletcher, Foorman, et al., 2002).

Learning disorders attributable to mental retardation, sensory problems (blindness, deafness), language status (e.g., English as a second language), or transient factors (adjustment difficulties, disruption of the home or school environment) should not be identified as LD. We have not included economic disadvantage, comorbid emotional and behavior disorders, or established neurological disorders as exclusionary criteria and would stipulate that the only way to exclude LD in children with these associated conditions is to provide an intervention that is appropriate and evaluate RTI. A classification of LD may exclude children with emotional or neurological disorders, or those who are economically disadvantaged from the LD category because of policy or resource issues—all are eligible for special education—but children with these associated conditions have forms of underachievement that are difficult to distinguish from those in children with LD. In the end, LD should be identified only after adequate opportunity to learn has been systematically evaluated. Those who do not respond to intervention need more specialized, individualized, and intensive treatments, as well as the probable conferment of disability status and the civil rights protections that come with identification. It is the intractability as indicated by an inadequate response to quality instruction that must be present to identify a child as LD. If a child responds, LD is

not indicated. Any child who has achievement difficulties should receive intervention, whether it is tutoring with support by a college student or intensive intervention by an experienced, well-trained academic therapist.

This model is quite different from the one that child clinical and other psychologists have utilized for the past few decades, and some may respond by suggesting that this model can only be implemented in schools. In fact, we argue that in the absence of an evaluation of RTI, LD should not be identified in any setting—school, clinic, hospital, and so on. We conceptualize traditional clinical evaluations as an opportunity to identify children as "at risk" for LD and to intervene with any child who is struggling to achieve. In schools, screening for reading problems can be done on a large-scale basis in kindergarten and Grade 1 as advocated in Donovan and Cross (2002) and implemented in states such as Texas (Fletcher, Foorman, et al., 2002). Those who are identified as at-risk have their progress monitored and receive increasingly intense, multitiered interventions that may eventuate in identification for special education in LD (Vaughn & Fuchs, 2003). In a multitiered intervention approach, children are screened for risk characteristics, such as weaknesses in letter-sound knowledge and phonological awareness in kindergarten and word reading in Grade 1, with immediate monitoring of progress (Torgesen, 2002). Depending on the rate of progress, interventions are intensified and modified in an effort to accelerate the rate of development of an academic skill. Children are not identified with disabilities until the final tier of the process.

Evaluations outside of schools should utilize a similar approach based on measurement of the three components of the hybrid model proposed by the consensus group in Bradley et al. (2002): (a) low achievement, (b) RTI, and (c) consideration of contextual factors and exclusions. Any psychological evaluation of a child or adolescent should consider the relevant achievement constructs (see following section) that represent the different types of LD. If there is evidence of low achievement, the focus should not be on extensive assessments of processing skills but on referral to an appropriate source for intervention. The psychologist should expect to have a working relationship with the intervention source so that RTI will be measured. This means that clinical child psychologists must be knowledgeable about educational interventions and prepared to develop a treatment plan that incorporates this form of intervention, just as they may be prepared to work with a physician around medication for problems with attention or anxiety. It is perfectly reasonable to ask the child to return every 4 to 6 months to repeat achievement tests and independently evaluate progress in conjunction with more frequent assessments of progress obtained by the intervention source.

The psychologist should also evaluate for other problems that may be associated with low achievement to adequately plan treatment. If mental retardation is suspected, IQ, adaptive behavior, and related assessments consistent with this classification can be administered. But note that if the child or adolescent has achievement scores in reading comprehension or math that are within two standard deviations of the mean (consistent with traditional legal definitions of mental retardation), or development of adaptive behavior obviously inconsistent with mental retardation, assessment of IQ is not necessary as such levels of performance preclude mental retardation. Some children may have oral language disorders that require speech and language intervention that will require referral and additional evaluation. Screening with vocabulary measures and through interacting with the child will help identify these children; the vocabulary screen will also help identify children who may benefit from additional intellectual screening. Many children with achievement difficulties or LD also have comorbid difficulties with attention and both internalizing and externalizing psychopathology. These disorders need to be assessed and treated, as simply referring a child for educational intervention without addressing comorbidities will surely increase the probability of a poor RTI. We believe that no clinical evaluation of a child should be conducted without a documentation of achievement levels through direct assessment or school report of such an assessment. If achievement deficits are apparent, intervention of some sort should be provided. It is not likely that treating a child for a comorbid disorder, such as ADHD, will result in improved levels of achievement in the absence of educational intervention.

Altogether, we are suggesting that from the perspective of LD, psychologists should perform assessments for emotional and behavioral disorders consistent with other articles in this special section. For LD, they need to administer achievement tests and evaluate RTI. This is regardless of subdiscipline (e.g., school psychologist, child clinical psychologist, neuropsychologist) or setting. To evaluate achievement, individualized norm-referenced assessments should be conducted. RTI requires assessments of intervention integrity and monitoring of progress.

Evaluating Achievement

Identifying specific achievement tests is not difficult, although tests for some domains are better developed than others. Lyon et al. (2003) suggested that LD represented six major achievement types, including (a) word recognition; (b) reading fluency; (c) reading comprehension; (d) mathematics computations; (e) reading and math, which is not really a comorbid association but a more severe reading problem with distinct math difficulties; and (f) written expression, which

could involve spelling, handwriting, or text generation. These patterns were drawn from the research literature (e.g., Rourke & Finlayson, 1978; Siegel & Ryan, 1988; Stothard & Hulme, 1996), but an extensive discussion of the evidence for these types is beyond the scope of this article (see Lyon et al., 2003). The assessment implications are straightforward. Many children and adolescents will have difficulties in more than one domain, so a thorough assessment of academic achievement is very important.

A set of achievement tests should be used. It is helpful to use tests from the same battery because the normative group is the same, which facilitates comparisons across tests. However, the battery from which these tests are chosen is less important than the constructs that are measured. Any single battery has strengths and weaknesses that can be supplemented with measures from other assessments. Given the suggestion that six types of LD may exist, the important constructs are word recognition, reading fluency, reading comprehension, math computations, and written expression. We usually assess spelling as a screen for written expression and handwriting difficulties and math and writing fluency as supplemental assessments.

Table 1 outlines these constructs and how they can be assessed with the commonly used Woodcock–Johnson Achievement Battery–III (WJ; Woodcock, McGrew, & Mather, 2001) or the Wechsler Individual Achievement Test–II (WIAT; Wechsler, 2001). We use the WJ and WIAT because they meet established criteria for reliability (internal consistency and test–retest) and validity (construct and concurrent) and were developed to account for variations in ethnicity and socioeconomic status. In particular, the normative sampling took into account this type of variation, and analyses (differential item functioning) were conducted to identify items that were not comparable across these sources of normative variation. There are also other norm-referenced assessments that can be used to supplement the WJ or WIAT, which we discuss later. Few of these supplemental measures have been developed with the care of the WJ or the WIAT, particularly with regard to the adequacy of the normative base and attempts to address different forms of normative variation.

Table 1 should not be taken to indicate that there are 11 different types of LD, one for each test. To reiterate, many children have problems in multiple domains. The pattern of academic strengths and weaknesses is an important consideration (Fletcher, Foorman, et al., 2002; Rourke, 1975). Table 1 identifies constructs and core tests that would be administered to every child and supplemental tests that would be used if there were concerns about a particular academic domain. If the referral indicated concerns about a particular area, additional tests from other measures would be used. Most children with significant academic problems where LD may eventually be a concern have difficulty with word recognition and consequently tend to have problems across domains of reading. Going beyond the core tests is usually not necessary if the child has problems with word recognition. Isolated problems with reading comprehension and written expression occur infrequently. If the problem is specifically math, using assessments in addition to the WJ or WIAT is helpful in ensuring that the deficiency is not just a matter of attention difficulties.

An advantage of the WJ and WIAT is the assessment of word recognition for both real words and pseudowords, the latter permitting an assessment of the child's ability to apply phonics rules to sound out words. Most achievement batteries assess recognition of real words, which is the essential component. These measures tend to be highly intercorrelated across different assessment batteries, including the Wide Range Achievement Test–III (Wilkinson, 1993), and the Accuracy measure from the Gray Oral Reading Test–Fourth Edition (Wiederholt & Bryant, 2001).

The WJ also has a silent reading speed subtest that, in our assessments, is highly correlated with other fluency measures despite the fact that it is not simply oral reading speed, requiring the child to answer some questions while reading a series of passages for 3 min.

Table 1. *Achievement Constructs in Relation to Subtests From the WJ and the WIAT*

Construct	WJ Subtest	WIAT Subtest
Core Tests		
Word Recognition	Word Identification	Word Reading
	Word Attack	Pseudoword Decoding
Reading Fluency	Reading Fluency	—
Reading Comprehension	Passage Comprehension	Reading Comprehension[a]
Math Computations	Calculation	Numerical Operations
Written Expression	Spelling	Spelling
Supplemental Tests		
Math Fluency	Math Fluency	—
Writing Fluency	Writing Fluency	—
Math Concepts	Quantitative Concepts	—
Written Expression	Writing Samples	Written Expression[a]

[a]Also assesses fluency

The WIAT permits assessment of reading speed during silent-reading comprehension. Both assessments are easily supplemented with the Test of Word Reading Efficiency (Torgesen, Wagner, & Raschotte 1999), which involves oral reading of real words and pseudowords on a list. The Test of Reading Fluency (Deno & Marston, 2001) is an option that requires text reading. Both measures are quick and efficient, and the former was designed with item analyses addressing differential item responses across ethnic groups. Whenever text is read out loud, fluency can be assessed as words read correctly per minute. The Gray Oral Reading Test–Fourth Edition includes a score for fluency of oral text reading.

Reading comprehension can only be screened with the WJ Passage Comprehension subtests which is a cloze-based assessment in which the child reads a sentence or passage and fills in a blank with a missing word. The Reading Vocabulary subtest is used to create a reading comprehension composite, but it places such a premium on decoding that we usually do not administer it. The WIAT also does not demand much reading of text. Some children who struggle to comprehend text in the classroom do not have difficulties on these subtests because the level of complexity rarely parallels what children are expected to read on an everyday basis. Supplemental assessments using the Group Reading Assessment and Diagnostic Education (Williams, Cassidy, & Samuels, 2001), the Gray Oral Reading Test–Fourth Edition, or even one of the well-constructed reading comprehension assessments from the group-based Stanford Achievement Test–10th Edition (Harcourt Assessment, 2002), Iowa Test of Basic Skills (Hoover, Hieronymous, Frisbie, & Dunbar, 2001), or similar instrument is essential. Often children have had these assessments in school, and it is useful to review results as part of the overall evaluation.

Reading comprehension is a difficult construct to assess (Francis, Fletcher, Catts, & Tomblin, in press). In evaluating comprehension skills, the assessor should attend to the nature of the material the child is asked to read and the response format. Reading comprehension tests vary in what the child reads (sentences, paragraphs, pages), the response format (cloze, open-ended questions, multiple-choice, think aloud), memory demands (answering questions with and without the text available), and how deeper aspects of meaning are evaluated (understanding of the essential meaning vs. literal understanding, vocabulary knowledge and elaboration, ability to infer or predict). It may be difficult to determine the source of the child's difficulties based on a single measure. Thus, if the issue is comprehension and the source is not in the child's word recognition or fluency skills, multiple measures that assess reading comprehension in different ways are needed.

For math, the Calculations subtest of the WJ and Numerical Operations subtest of the WIAT are paper-and-pencil tests of math computations (Table 1). Low scores on this type of task predict variation in cognitive skills depending on other academic strengths and weaknesses (Rourke, 1993). However, low scores could reflect problems with fact retrieval and verbal working memory if word recognition is comparably lower, as opposed to problems with procedural knowledge if word recognition is significantly higher and not deficient. Deficient scores can also reflect problems paying attention, especially in children with ADHD. The math computations subtests from the Wide Range Achievement Test–III is also frequently used and is useful because it is timed and the problems are less organized. The key is the paper-and-pencil assessment of math computations, which is how difficulties in math are typically manifested in children who do not have reading problems. As in reading, assessments of fluency are helpful, although there is no evidence suggestive of a math fluency disorder. In Table 1, the WJ Math Fluency subtest is identified as a supplemental measure, representing a timed assessment of single-digit arithmetic facts that may be helpful for identifying children who lack speed in basic arithmetic skills. Such difficulties make it difficult to master more advanced aspects of mathematics. If an assessment of math concepts is needed, which we would do only if math was an overriding concern, the Quantitative Concepts subtest of the WJ is more useful than the WJ Applied Problems or WIAT Math Reasoning subtests, which introduce word problems that are difficult for children with reading difficulties.

Written expression is most difficult to assess, partly because it is not clear what constitutes a disorder of written expression—spelling, handwriting, or text generation (Lyon et al., 2003). Obviously problems with the first two components will constrain text generation. Spelling should be assessed as it may represent the primary source of difficulty with written expression for children, especially if they also have word-recognition difficulties. The analysis of spelling errors (Rourke, Fisk, & Strang, 1986) can be informative in understanding whether the problem is with the phonological component of language or with the visual form of letters (i.e., orthography). Spelling also permits an informal assessment of handwriting.

Table 1 identifies WJ and WIAT measures of written expression. The utility of these measures is not well established, and the significant generation of text in terms of construction and writing of passages and stories is not really required. As with reading comprehension, it may be important to supplement or even replace this assessment with a test such as the Thematic Maturity subtest of the Test of Written Language (Hammill & Larsen, 1998). Measuring fluency with a measure such as the WJ Writing Fluency subtest may also be

useful. As in reading and math, fluency of writing predicts the quality of composition.

From this type of assessment, characteristic patterns emerge that will demarcate the classification and indicate a need for specific kinds of intervention. For each of the six types of LD, there are interventions with evidence of efficacy that should be utilized in or out of a school setting (Lyon et al., in press). The goal is not to diagnose LD, which is not feasible in a one-shot evaluation for the psychometric and conceptual reasons outlined previously, but to identify achievement difficulties that can be addressed through intervention. If the assessor is knowledgeable about these patterns, very specific intervention recommendations, as well as the need for other assessments, can be made.

Table 2 summarizes achievement patterns that are well established in research (Fletcher, Foorman, et al., 2002; Lyon et al., 2003). Intervention should be considered for any child who performs below the 25th percentile on a well-established assessment, with an understanding that these are not firm cut points and should be evaluated across all the measures. We are not indicating that 25% of all children have a LD, only that scores below the 25th percentile are commonly associated with low performance in school, assuming the cut point is reliably attained. In examining Table 2, the decision rules should not be rigidly applied and are simply guidelines to assist clinicians. Identifying LD is always based on factors beyond just the test scores. The decision process should focus on what is needed for intervention, which requires an assessment of contextual variables and the presence of comorbid disorders that influence decisions about what sort of plan will be most effective for an individual child. Low achievement is related to many contextual variables, which is why the flexibility in special education guidelines allows interdisciplinary teams to base decisions on factors that go beyond test scores. The purpose of assessment is ultimately to develop an intervention plan.

Evaluating Response to Instruction

Once a child is screened or tested for achievement deficits, progress should be monitored if a problem is identified. It is astonishing that U.S. special education guidelines do not require at least yearly readministration of the achievement tests that were used to justify the placement as one method of assessing the efficacy of the intervention plan. If a child is responding to intervention, his or her rate of development should be accelerated relative to the normative population (i.e., the achievement gap is closed). As part of this assessment of RTI, progress should be monitored on a frequent basis if the problem is with word recognition or fluency, math computations, or spelling. Reading comprehension and higher forms of written expression will show less rapid change and progress, as monitoring tools for these types of problems have not been adequately developed.

Most of the tests mentioned here have alternative forms. But some have been developed to permit assessments with even more frequency and are referred to "curriculum-based assessments" (Fuchs & Fuchs,

Table 2. *Eight Achievement Patterns Associated With Intervention*

1. Decoding and Spelling < 90; Arithmetic one half standard deviation higher than word recognition and spelling and at least 90. This is a pattern characterized by problems with single word decoding skills and better arithmetic ability. Reading comprehension will vary depending on how it is assessed but is usually impaired. Children with this pattern have significant phonological language problems and strengths in spatial and motor skills (Rourke & Finlayson, 1978).
2. Arithmetic < 90, Decoding and Spelling > 90 and at least 7 points higher. Children with difficulties that only involve math show this pattern, which is associated with problems with motor and spatial skills, problem-solving deficiencies, and disorganization (Rourke & Finlayson, 1978). It usually represents problems with math procedures as opposed to math facts (Lyon et al., 2003).
3. Decoding, Comprehension, Spelling, and Arithmetic < 90. This pattern represents a problem with word recognition characterized by language and working memory problems more severe than in children with poor decoding and better development of math skills (Rourke & Finlayson, 1978). The math problem involves learning and retrieving math facts (Lyon et al., 2003).
4. Spelling and Arithmetic < 90, Decoding > 90 and 7 points higher. Essentially the same pattern as Number 3 except the motor (and writing) component is more severe.
5. Reading Comprehension < 90 and 7 points below decoding. This pattern often reflects long-term oral language disorder. Problems with receptive language, short-term memory, and attention are apparent, with strengths in phonological language skills (Stothard & Hulme, 1996).
6. Decoding skills 7 points lower than Comprehension skills and < 90. This pattern reflects a phonological language problem with usually better than average semantic language and spatial skills (Stothard & Hulme, 1996). The pattern is not apparent for reading comprehension measures that are timed or require significant amounts of text reading.
7. Reading Fluency < 90 and < Decoding by one half standard deviation will reflect a problem where accuracy of word reading is less of a problem than automaticity of word reading (Lyon et al., 2003).
8. Spelling < 90. This pattern reflects (a) motor deficits in a young child or (b) residuals of earlier phonological language problems that have been remediated or compensated in older children and adults. The pattern is common in adults with a history of word recognition difficulties. Fluency is often impaired.

Note: The patterns are based on relations of reading decoding, reading fluency, reading comprehension, spelling, and arithmetic. It is assumed that any score below the 25th percentile (standard score = 90) is impaired and that a difference of one half standard deviations is important (± 7 standard score points). The patterns should be considered prototypes and the rules loosely applied (adapted from Fletcher, Foorman, Boudousquie, Barnes, Schatschneider, & Francis, 2002). These patterns are not related to IQ scores.

1999). Such measures are often used by the intervener (e.g., teacher) to document how well a child is responding to instruction. Typically a child would read a short reading passage appropriate for grade level (or do a set of math computations) for 2 to 3 min. The number of words (or math problems) correctly read (or computed) would be graphed over time and compared against grade-level benchmarks, representing a criterion-referenced form of assessment. A child may be screened with these measures, and those performing below the benchmark may be candidates for intervention, especially in schools.

Such assessments should also be accompanied by observations of the integrity of the implementation of the intervention, including the amount of time spent on supplemental instruction, especially if the child does not appear to be making progress. School psychologists are often well prepared in this area of assessment. Although a psychologist operating outside of a school may not be in a position to do curriculum-based assessments or to personally evaluate the intervention, such assessments should be expected, especially if the referral is to a private academic therapist.

A variety of methods have been developed, and the assessments with the most widespread utilization are the Monitoring Basic Skills Progress (Fuchs, Hamlett, & Fuchs, 1990), which assesses reading, math, and spelling fluency, and the Dynamic Indicators of Basic Early Reading Skills (Good, Simmons, & Kame'enui, 2001), a battery of different reading fluency measures. Some of these tools are focused primarily on the lower grades, but the norm-referenced assessments of fluency identified previously—especially if they have alternative forms—can be used with older students,. These measures meet accepted psychometric criteria for reliability and validity. The curriculum-based assessment measures have not been assessed as formally for differential item functioning but have been widely employed with school populations that are quite diverse (Fuchs & Fuchs, 1999; Shinn, 1998).

Conclusions

Based on our evaluation of models, we propose a hybrid model that incorporates features of low achievement and RTI models for the identification of children as LD. We specifically do not find evidence to support extensive assessments of cognitive, neuropsychological, or intellectual skills to identify children as LD. Although some may view this model as only for schools, we reject the idea that the routine evaluations done in the past by psychologists and educators outside of school settings are useful for LD. We find little value in the idea of evaluating a child in a single assessment and concluding that the child has LD based on an IQ-achievement discrepancy, low achievement, or profiles on neuropsychological tests, largely because such assessments are not directly related to treatment and the diagnosis itself is not reliable. As soon as it is apparent that the child has an achievement problem, a referral for intervention should be made and the resources that might be spent on diagnosis should be spent on intervention. Children should not be diagnosed as LD until a proper attempt at instruction has been made. Assessment of achievement skills should be a routine part of any psychological evaluation of a child and cannot be seen as the province of just the schools. Serial monitoring of RTI and the integrity of instruction should be completed before children are identified as LD. There are issues involved in the intervention component, estimation of slope and intercept effects, as well as decisions that have to be made about cut points that will differentiate responders and nonresponders (Gresham, 2002). For these reasons alone, RTI cannot be the sole criterion for identification, and flexibility in decision making is required. At the same time, there appears to be considerable validity to this approach, implying that it is indeed possible to reliably identify nonresponders as a group with unexpected underachievement.

In addition to the evidence for validity (and the greater reliability of the underlying psychometric model), the model does not require the use of exclusionary criteria (especially emotional disturbance and economic disadvantage) to operationalize unexpected underachievement, thus capturing the construct of LD. This is an important consideration given the lack of evidence validating classifications that utilize these particular exclusions (Kavale, 1988; Lyon et al., 2001). The model does operationalize the concept of opportunity to learn, which is rarely directly assessed as part of LD identification. It is also a model that can only be implemented in an instructional setting, such as a school, or in clinical settings outside of public schools where remediation is utilized, such as an academic therapy setting. But it is not consistent with the traditional approach to LD identification based on a single administration of a test battery and consideration of a diagnosis, which we believe is an outmoded model that detracts from intervention. In the absence of an attempt to systematically instruct the child, LD cannot be "diagnosed," obviating the traditional "test and treat" model, as identifying LD must be the end product of an attempt to instruct the child (i.e., "treat and test"). This is not a post hoc approach but rather an argument that in the absence of the opportunity to learn exclusion, the concept of LD has no basis in evidence, and low achievement per se is not adequate evidence for LD. Such an approach ties the concept of LD to treatment, which is important. It may be that a single assessment may indicate "risk" or even an achievement disorder. But such an assessment cannot indicate a "disability" in the absence of functional criteria that would include opportunity to learn.

A final comment involves what some will see as equating LD with measurable deficits on achievement tests. Some will argue that the mere presence of a deficit on a measure of processing skills means that person should be identified by LD, in part because of the belief that such deficits indicate a brain anomaly. The most common example is the linking of "executive function" deficits with LD. We argue that the concept of LD is empty without a focus on achievement, largely because it becomes more difficult to identify a unique subgroup representing LD that would be different from other classes of childhood disorders. Executive functions, for example, are often linked to ADHD, but classifications of ADHD based on executive function deficits as assessed by cognitive tests do not have much validity (Barkley, 1997). Moreover, executive function deficits characterize many childhood populations.

More fundamentally, consider an overarching classification of childhood learning and behavioral difficulties. For LD, achievement deficits represent markers for the underlying classification. What distinguishes the LD prototype from, for example, a behavioral disorder such as ADHD is the presence of an achievement deficit. If a child with ADHD has an achievement deficit, it is usually reflective of a comorbid association (Fletcher, Shaywitz, & Shaywitz, 1999). If we expand our classification to mental retardation, the key for differentiating mental retardation from LD (or ADHD) is not just the intelligence test score. Rather, the major difference is in adaptive behavior, where mental retardation should reflect a pervasive deficit in adaptive behavior and LD as a relatively narrow deficit (Bradley et al., 2002). So a classification of these three major disorders requires markers for achievement, attention-related behaviors, and adaptive behavior. In the absence of these types of markers, and a focus on classification, all children with problems are simply disordered and there is no need for assessment because they would all require the same interventions. When LD is tied to levels and patterns of achievement, an evidence base for differential interventions focused on learning in specific academic domains emerges. This is the strongest evidence for the validity of the concept of LD, its classification, and the source of evidence-based approaches to assessment and identification.

References

Aaron, P. G., Kuchta, S., & Grapenthin, C. T. (1988). Is there a thing called dyslexia? *Annals of Dyslexia, 38,* 33–49.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkley, R. A. (1997). *ADHD and the nature of self-control.* New York: Guilford.

Bereiter, C. (1967). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in the measurement of change.* Madison: University of Wisconsin Press.

Bradley, R., Danielson, L., & Hallahan, D. P. (Eds.). (2002). *Identification of learning disabilities: Research to practice.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christensen, C. A. (1992). Discrepancy definitions of reading disability: Has the quest led us astray? *Reading Research Quarterly, 27,* 276–278.

Cisek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cisek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–18). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Deno, S. L., & Marston, D. (2001). *Test of Oral Reading Fluency.* Minneapolis: Educators Testing Service.

Donavon, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

Fisher, S. E., & DeFries, J. C. (2002). Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nature: Review of Neuroscience, 10,* 767–780.

Fletcher, J. M., Foorman, B. R., Boudousquie, A. B., Barnes, M. A., Schatschneider, C., & Francis, D. J. (2002). Assessment of reading and learning disabilities: A research-based, intervention-oriented approach. *Journal of School Psychology, 40,* 27–63.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, B. A., & Shaywitz, S. E. (1993). Classification of learning disabilities: Relationships with other childhood disorders. In G. R. Lyon, D. Gray, J. Kavanagh, & N. Krasnegor (Eds.), *Better understanding learning disabilities* (pp. 27–55). New York: Brookes.

Fletcher, J. M., Lyon, G. R., Barnes, M. Stuebing, K. K., Francis, D. J., Olson, R., et al. (2002). Classification of learning disabilities: An evidence-based evaluation. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 185–250). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fletcher, J. M., Morris, R. D., & Lyon, G. R. (2003). Classification and definition of learning disabilities: An integrative perspective. In H. L. Swanson, K. R. Harris, & S. Graham. *Handbook of learning disabilities* (pp. 30–56). New York: Guilford.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., et al. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 85,* 1–18.

Fletcher, J. M., Shaywitz, S. E., & Shaywitz, B. A. (1999). Comorbidity of learning and attention disorders: Separate but equal. *Pediatric Clinics of North America, 46,* 885–897.

Fletcher, J. M., Simos, P. G., Papanicolaou, A. C., & Denton, C. (2004). Neuroimaging in reading research. In N. K. Duke & M. H. Mallette (Eds.), *Literacy research methods* (pp. 252–286). New York: Guilford.

Francis, D. J., Fletcher, J. M., Catts, H., & Tomblin, B. (in press). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Current issues in reading comprehension and assessment.* Mahwah, NJ: Lawrence Erlbaum Associate, Inc.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of learning disabilities: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38,* 98–108.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88,* 3–17.

Fuchs, L., & Fuchs, D. (1998). Treatment validity: A simplifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice, 4,* 204–219.

Fuchs, L., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *The School Psychology Review, 28,* 659–671.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Monitoring basic skills progress.* Austin, TX: PRO-ED.

Gilger, J. W. (2002). Current issues in the neurology and genetics of learning-related traits and disorders: Introduction to the special issue. *Journal of Learning Disabilities, 34,* 490–491.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257–288.

Gresham, F. M. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467–564). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language.* Austin, TX: PRO-ED.

Harcourt Assessment. (2002). *Stanford Achievement Test* (10th ed.). New York: Author.

Hoover, A. N., Hieronymous, A. N., Frisbie, D. A., & Dunbar, S. B. (2001). *Iowa Test of Basic Skill.* Itasca, IL: Riverside.

Hoskyn, M., & Swanson, H. L (2000). Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. *The School Psychology Review, 29,* 102–119.

Individuals With Disabilities Education Act, 20 U. S. C. 1400 *et seq.* Stat. 34 C. F. R. 300 (1997).

Jiménez, J. E., del Rosario Ortiz, M., Rodrigo, M., Hernández-Valle, I., Ramirez, G., Estévez, A., et al. (2003). Do the effects of computer-assisted practice differ for children with reading disabilities with and without IQ-achievement discrepancy? *Journal of Learning Disabilities, 36,* 34–47.

Jorm, A. F., Share, D. L., Matthews, M., & Matthews, R. (1986). Cognitive factors at school entry predictive of specific reading retardation and general reading backwardness: A research note. *Journal of Child Psychology, 27,* 45–54.

Kavale, K. A. (1988). Learning disability and cultural disadvantage: The case for a relationship. *Learning Disability Quarterly, 11,* 195–210.

Lewis, C., Hitch, G. J., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology Psychiatry, 35,* 283–292.

Lyon, G. R., Fletcher, J. M., & Barnes, M. C. (2003). Learning disabilities. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (2nd ed.., pp. 520–586). New York: Guilford.

Lyon, G. R., Fletcher, J. M., Fuchs, L., & Chhabra, V. (in press). Treatment of learning disabilities. In E. Mash & R. Barkley (Eds.), *Treatment of childhood disorders.* New York: Guilford.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., et al. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., R. A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute.

Lyon, G. R., & Moats, L. C. (1988). Critical issues in the instruction of the learning disabled. *Journal of Consulting and Clinical Psychology, 56,* 830–835.

MacMillan, D. L., & Siperstein, G. N. (2002). Learning disabilities as operationally defined by schools. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 287–368). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mazzocco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school age years. *Annals of Dyslexia, 53,* 218–253.

Mercer, C. D., Jordan, L., Allsop, D. H., & Mercer, A. R. (1996). Learning disabilities definitions and criteria used by state education departments. *Learning Disability Quarterly, 19,* 217–232.

Morris, R., & Fletcher, J. M. (1988). Classification in neuropsychology: A theoretical framework and research paradigm. *Journal of Clinical and Experimental Neuropsychology, 10,* 640–658.

Morris, R. D., Fletcher, J. M., & Francis, D. J. (1993). Conceptual and psychometric issues in the neuropsychological assessment of children: Measurement of ability discrepancy and change. In I. Rapin & S. Segalovitz (Eds.), *Handbook of neuropsychology* (Vol. 7, pp. 341–352). Amsterdam: Elsevier.

Morris, R. D., Stuebing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., et al. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology, 90,* 347–373.

National Center for Learning Disabilities. (2002). *Achieving better outcomes—maintaining rights: An approach to identifying and serving students with specific learning disabilities.* New York: Author.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: National Institute of Child Health and Human Development.

Pennington, B. F., Gilger, J. W., Olson, R. K., & DeFries, J. C. (1992). The external validity of age- versus IQ-discrepancy definitions of reading disability: Lessons from a twin study. *Journal of Learning Disabilities, 25,* 562–573.

President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families.* Washington, DC: Department of Education.

Reschly, D. J., Hosp, J. L., & Smied, C. M. (2003). *And miles to go ...: State SLD requirements and authoritative recommendations.* National Research Center on Learning Disabilities. Retrieved July 22, 2003, from the World Wide Web: www.nrcld.org

Reschly, D. J., Tilly, W. D., & Grimes, J. P. (1999). *Special education in transition: Functional assessment and noncategorical programming.* Longmont, CO: Sopris West.

Reynolds, C. (1984–1985). Critical measurement issues in learning disabilities. *Journal of Special Education, 18,* 451–476.

Rodgers, B. (1983). The identification and prevalence of specific reading retardation. *British Journal of Educational Psychology, 53,* 369–373.

Rogosa, D. (1995). Myths about longitudinal research (plus supplemental questions). In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rourke, B. P. (1975). Brain-behavior relationships in children with learning disabilities: A research program. *American Psychologist, 30,* 911–920.

Rourke, B. P. (1993). Arithmetic disabilities specific or otherwise: A neuropsychological perspective. *Journal of Learning Disabilities, 26,* 214–226.

Rourke, B. P., & Finlayson, M. A. J. (1978). Neuropsychological significance of variations in patterns of academic performance: Verbal and visual-spatial abilities. *Journal of Pediatric Psychology, 3,* 62–66.

Rourke, B. P., Fisk, J., & Strang, J. (1986). *Neuropsychological assessment of children.* New York: Guilford.

Shalev, R. S., Auerbach, J., Manor, O., & Gross-Tsur, V. (2000). Developmental dyscalculia: Prevalence and prognosis. *European Adolescent Psychiatry, 9,* 58–64.

Shaywitz, S. E., Escobar, M. D., Shaywitz, B. A., Fletcher, J. M., & Makuch, R. (1992). Distribution and temporal stability of dyslexia in an epidemiological sample of 414 children followed longitudinally. *New England Journal of Medicine, 326,* 145–150.

Shaywitz, S. E., Shaywitz, B. A., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., et al. (2003). Neural systems for compensation and persistence: Young adult outcomes of childhood reading disability. *Biological Psychiatry, 54,* 25–33.

Shepard, L. (1980). An evaluation of the regression discrepancy method for identifying children with learning disabilities. *Journal of Special Education, 14,* 79–91.

Shinn, M. R. (1998). *Advanced applications of curriculum-based measurement.* New York: Guilford.

Siegel, L. S. (1992). An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities, 25,* 618–629.

Siegel, L. S., & Ryan, E. (1988). Working memory in subtypes of learning disabled children. *Journal of Clinical and Experimental Neuropsychology, 10,* 55.

Silva, P. A., McGee, R., & Williams, S. (1985). Some characteristics of 9-year-old boys with general reading backwardness or specific reading retardation. *Journal of Child Psychology and Psychiatry, 26,* 407–421.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identify early reading disability. *Journal of Educational Psychology, 93,* 735–749.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36,* 24–33.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86,* 24–53.

Stothard, S. E., & Hulme, C. (1996). A comparison of reading comprehension and decoding difficulties in children. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 93–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518.

Torgesen, J. K. (2002). Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 565–650). Mahwah, NJ: Lawrence Erlbaum. Associates, Inc.

Torgesen, J., Wagner, R., & Raschotte, C. (1999). *Test of Word Reading Efficiency.* Austin, TX: PRO-ED.

U. S. Office of Education. (1977). Assistance to states for education for handicapped children: Procedures for evaluating specific learning disabilities. *Federal Register, 42,* G1082–G1085.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction. *Learning Disabilities Research and Practice, 18,* 137–146.

Vaughn, S. R., Linan-Thompson, S., & Hickman-Davis, P. (2003). Response to treatment as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391–409.

Vellutino, F. R. (1979). *Dyslexia: Theory and research.* Cambridge, MA: MIT Press.

Vellutino, F. R., Fletcher, J. M., Scanlon, D. M., & Snowling, M. J. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychiatry and Psychology, 45,* 2–40.

Vellutino, F. R., Scanlon, D. M., & Jaccard, J. (2003). Toward distinguishing between cognitive and experiential deficits as primary sources of difficulty in learning to read: A two year follow-up of difficult remediate and readily remediate poor readers. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 73–120). Baltimore: York.

Wadsworth, S. J., Olson, R. K., Pennington, B. F., & DeFries, J. C. (2000). Differential genetic etiology of reading disability as a function of IQ. *Journal of Learning Disabilities, 33,* 192–199.

Wechsler, D. (2001). *Wechsler Individual Achievement Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Wiederholt, J. L, & Bryant, B. R. (2001). *Gray Oral Reading Test* (4th ed.). Austin, TX: PRO-ED.

Wilkinson, G. (1993). *Wide Range Achievement Test–3.* Wilmington, DE: Wide Range.

Williams, K. T., Cassidy, J., & Samuels, S. J. (2001). *Group reading assessment and diagnostic education.* Circle Pines, MN: American Guidance Services.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III.* Itasca, IL: Riverside.

Received March 4, 2004
Accepted November 10, 2004

SOCIAL BEHAVIOR AND PERSONALITY, 2014, 42(4), 625-634

http://dx.doi.org/10.2224/sbp.2014.42.4.625

THE RELATIONSHIPS AMONG STERNBERG'S TRIARCHIC ABILITIES, GARDNER'S MULTIPLE INTELLIGENCES, AND ACADEMIC ACHIEVEMENT

BIRSEN EKINCI
Marmara University

In this study I investigated the relationships among Sternberg's Triarchic Abilities (STA), Gardner's multiple intelligences, and the academic achievement of children attending primary schools in Istanbul, Turkey. Participants were 172 children (93 boys and 81 girls) aged between 11 and 12 years. STA Test (STAT) total scores were significantly and positively related to linguistic, logical-mathematical, and intrapersonal test scores. Analytical ability scores were significantly positively related to only logical-mathematical test scores, practical ability scores were only related to intrapersonal test scores, and the STAT subsections were significantly related to each other. After removing the effect of multiple intelligences, the partial correlations between mathematics, social science, and foreign language course grades and creative, practical, analytical, and total STAT scores, were found to be significant for creative scores and total STAT scores, but nonsignificant for practical scores and analytical STAT scores.

Keywords: Sternberg's Triarchic Abilities Test, multiple intelligences, academic achievement, children, intelligence.

Since 1980 there has been increasing interest in the role of intelligence in learning and its impact on student achievement. Similarly to education theorists, many researchers on intelligence have been conducting studies to apply theories about intelligence, to education in general and, in particular, to the instructional context of the classroom (Castejón, Gilar, & Perez, 2008). The main difference between contemporary and older approaches to the role of intelligence is that,

Birsen Ekinci, Atatürk Education Faculty, Marmara University.
This study was supported by the Marmara University, Scientific Research Projects Center, research number EGT-D-110913-0387.
Correspondence concerning this article should be addressed to: Birsen Ekinci, Atatürk Education Faculty, Department of Primary Education, Marmara University, Göztepe Campus, 34722 Kadiköy, Istanbul, Turkey. Email: **birsenp@marmara.edu.tr**

in earlier conceptualizations, intelligence was described as involving one factor of general mental ability that encompasses the common variance among all the contributing factors. The existence of this general intelligence factor was originally hypothesized by Spearman in 1927 and labeled as "g" (see Jensen, 1998). It was hypothesized that this g factor exists over and above the various abilities that make up intelligence, including verbal, spatial visualization, numerical reasoning, mechanical reasoning, and memory (Carroll, 1993). However, according to contemporary theories, intelligence must be regarded as existing in various forms and the levels of intelligence can be improved through education. The most widely accepted comparative theories of intelligences in recent literature are Gardner's (1993) multiple intelligences theory and Sternberg's (1985) triarchic theory of intelligence. Researchers have reported significant differences between student outcomes for classroom instruction conducted following the principles of multiple intelligences, and student outcomes under traditionally designed courses of instruction in science (Özdermir, Güneysu, & Tekkaya, 2006), reading (Al-Balhan, 2006), and mathematics (Douglas, Burton, & Reese-Durham, 2008).

Gardner (1993) developed a theory of multiple intelligences that comprises seven distinct areas of skills that each person possesses to different degrees. Linguistic intelligence (LI) is the capacity to use words effectively, either orally or in writing. Logical-mathematical intelligence (LMI) is the capacity to use numbers effectively and to reason well. Spatial intelligence (SI) is the ability to perceive the visual-spatial world accurately and to interpret these perceptions. Bodily-kinesthetic intelligence (KI) involves expertise in using one's body to express ideas and feelings. Musical intelligence (MI) is the capacity to perceive, discriminate, and express musical forms. Interpersonal intelligence (INPI) is the ability to perceive, and make distinctions in, the moods, intentions, motivations, and feelings of other people. Intrapersonal intelligence (INTI) is self-knowledge and the ability to act adaptively on the basis of that knowledge. Naturalist intelligence (NI) is expertise in the recognition and classification of the numerous species – the flora and fauna – of a person's environment (Armstrong, 2009).

Researchers have addressed the relationship between multiple intelligences and metrics of different abilities, and of various psychological constructs. Reid, Romanoff, Algozzine, and Udall (2000) showed that SI, LI, and LMI were related to scores in a test to measure the nonverbal abilities of pattern completion, reasoning by analogy, serial reasoning, and spatial visualization, among a group of handicapped and nonhandicapped children aged between 5 and 17 years. Furthermore, the effects of multiple intelligences-based teaching strategies on students' academic achievement have been studied extensively (Al-Balhan, 2006; Douglas et al., 2008; Greenhawk, 1997; Mettetal, Jordan, & Harper, 1997; Özdermir et al., 2006). In addition, some researchers have investigated the relationship between multiple intelligences and academic achievement (McMahon, Rose, & Parks, 2004; Snyder, 1999). McMahon and colleagues

found that, compared with other students, fourth-grade students with higher scores on LMI were more likely to demonstrate reading comprehension scores at, or above, grade level. In a similar study, Snyder reported a positive correlation between high school students' grade point averages and KI. In the same study results showed that there was a positive correlation between the total score for the Metropolitan Achievement Test-Reading developed by the Psychological Corporation of San Antonio, Texas, USA and the categories of LMI and LI.

Sternberg developed the second well-known intelligence theory. According to Sternberg (1999a, 1999b), individuals show their intelligence when they apply the information-processing components of intelligence to cope with relatively novel tasks and situations. Within this approach to intelligence, Sternberg (1985) proposed the triarchic theory of intelligence, according to which there are three different, but interrelated, aspects of intellect: (a) analytic intelligence, (b) creative intelligence, and (c) practical intelligence. Individuals highly skilled in analytical intelligence are adept at analytical thinking, which involves applying the components of thinking to abstract, and often academic, problems. Individuals who have a high degree of creative intelligence are skilled at discovering, creating, and inventing ideas and products. People who have a high level of practical intelligence are good at using, implementing, and applying ideas and products. Sternberg (1997) developed an instrument, the Sternberg Triarchic Abilities Test (STAT), to evaluate triarchically based intelligence. In this instrument each aspect of intelligence is tested through three modes of presentation of problems: verbal, quantitative, and figural. A number of previous researchers have established the construct validity of the STAT (Sternberg, Castejón, Prieto, Hautamäki, & Grigorenko, 2001; Sternberg, Ferrari, Clinkenbeard, & Grigorenko, 1996). Although Sternberg did not intend the STAT to be a measure of general intelligence, as assessed by conventional intelligence tests, in related literature (Brody, 2003) there are contradictory results and opinions on this issue. Sternberg (2000a, 2000b) has claimed that the STAT is independent of measures of general intelligence and a more accurate predictor of academic achievement. However, Gottfredson (2002) pointed out that the data obtained to support this claim are sparse and suggested that the data collected by Sternberg et al. (1996) support the conclusion that the STAT is related to other measures of intelligence and may, in fact, be a measure of general intelligence. The triarchic abilities are related to different intelligence tests scores (e.g., Concept Mastery Test, Watson Glaser Critical Thinking Appraisal, Cattle Culture-Fait Test of g; Sternberg et al., 1996). However, Brody (2003) suggested that although these correlations are substantial, it is likely that they underestimate general intelligence because they were obtained from a sample of high school students who were predominately categorized as gifted, as determined by IQ scores, and these students were, therefore, likely to record a restricted range of scores on the tests.

In the present study I hypothesized that both multiple intelligences total scores and STAT total scores would be predictors of academic achievement. Specifically, I hypothesized that the LI and LMI, and the analytical STAT, would be predictors of student success in the subject areas of mathematics, science, social science, and foreign- language learning.

Method

Participants

Participants were 174 randomly selected fifth- and sixth-grade students (81 girls and 93 boys) attending primary school in Istanbul, Turkey. Students' ages ranged from 11 to 12 years old.

Instruments

The students completed the Turkish version of Gardner's Multiple Intelligences Inventory (MII; Saban, 2002) to assess participants' preferred intelligence within one of the eight categories: LI, LMI, SI, MI, KI, INPI, INTI, and NI. The possible score for the MII ranges from 0 to 80. The individual category in which a student has the highest score is considered to be the type of intelligence in which that student is most skilled. The overall Cronbach's alpha reliability coefficient in this study was .96, denoting high reliability; .89 for LI; .83 for LMI; .89 for SI; .88 for MI; .78 for KI; .85 for INPI; .85 for INTI; and .84 for NI.

The second instrument that I used in this study was Sternberg's Triarchic Abilities Test (STAT). The test comprises 81 items divided across three subsections designed to measure analytical, creative, and practical abilities. I translated this test into Turkish using the back-translation technique. In order to ensure that the back-translation retained the meaning of the original form, I conducted validity and reliability checks. The Turkish and the English versions of the test were given to 80 bilingual Turkish- and English-speaking students to complete within two weeks. Analyses of scores for the Turkish and English versions of test completed by these students yielded high correlation values (.85 for analytical, .79 for practical, and .81 for creative subsections). The overall alpha reliability coefficient of this test was .89, and for the subsections it was .80 for analytical, .77 for practical, and .78 for creative.

Procedure

The students completed the instruments during class time and in their classrooms. There was no time limit for completion. Each test session lasted approximately 60 minutes. The parents of the participating children gave permission for the researcher to access the students' grade point average for mathematics, science, social science, and foreign language courses at the end of the year during which the study was conducted. Each participant received a pen and pencil as a thank-you gift for his/her participation in this study.

Data Analysis

The data were analyzed using SPSS version 15 to conduct correlation analysis and multiple regression analysis.

Results

As shown in Table 1, the children's STAT total scores (M = 35.34, SD = 9.09) were significantly and positively related to LI (M = 28.98; SD = 7.59), LMI (M = 30.12, SD = 6.87), and INTI (M = 29.10, SD = 7.15) scores ($p < .01$). Analytical subsection STAT scores (M = 13.76, SD = 3.96) were significantly related to LM intelligence scores ($p < .01$). STAT practical subsection scores (M = 10.37, SD = 3.06) were significantly correlated only with INTI scores ($p < .01$).

Table 1. *Relationships Among STAT Total Scores, Analytical, Practical, and Creative Ability Scores, and Multiple Intelligences Scores*

	LI	LMI	SI	MI	KI	INPI	INTI	NI
Analytical	.303	.413**	-.057	.093	.036	.021	.281	-.102
Practical	.274	.268	.003	.113	.041	.095	.434**	-.109
Creative	.291	.540**	-.062	.103	.004	-.049	.361*	-.098
Total	.351*	.506**	-.051	.123	.031	.019	.425**	-.124

Note. ** $p < .01$, * $p < .05$. LI = linguistic intelligence, LMI = logical-mathematical intelligence, SI = spatial intelligence, MI = musical intelligence, KI = bodily-kinesthetic intelligence, INPI = interpersonal intelligence, INTI = intrapersonal intelligence, NI = naturalist intelligence.

Mathematics course grades (M = 3.78; SD = 1.20) were significantly related to the STAT total ($p < .001$) and to the STAT analytical ($p < .001$), practical ($p < .01$), and creative ($p < .01$) subsections. Similarly, social science (M = 3.78, SD = 1.10) and science course grades (M = 3.51, SD = 1.40) were significantly related to the STAT total ($p < .01$) and to the STAT analytical ($p < .01$) and creative ($p < .01$) subsections. However, foreign language course grades (M = 3.57, SD = 1.16) were significantly related to all of the subsection scores of the STAT ($p < .001$; see Table 2).

Table 2. *Relationships Among STAT Total Scores, Analytical, Practical, and Creative Subsection Scores, and Academic Success*

	Mathematics	Science	Social science	Foreign language
Analytical	.536*	.395**	.304**	.454*
Practical	.461**	.264	.269	.451*
Creative	.491*	.378**	.307**	.442*
Total	.588*	.415**	.347**	.527*

Note. * $p < .001$, ** $p < .01$.

Mathematics grades of the participants were significantly related to LI ($p < .01$), LMI ($p < .01$), INPI ($p < .05$), and INTI ($p < .01$) scores. Similarly, students' course grades for science were significantly related to LI ($p < .05$), LMI ($p < .01$), and INTI ($p < .05$) scores; students' social science course grades were significantly related to LI ($p < .05$), LMI ($p < .01$), and INTI ($p < .05$) scores; and students' course grades for foreign languages were significantly related to LI ($p < .01$), LMI ($p < .01$) and INTI ($p < .01$) scores (see Table 3).

Table 3. *Relationships Between Multiple Intelligences Scores and Academic Success*

	LI	LMI	SI	MI	KI	INPI	INTI	NI
Mathematics	.458**	.695**	.080	.174	.285	.356*	.522**	.140
Science	.340*	.575**	.007	.070	.239	.312	.379*	.085
Social science	.359*	.598**	.125	.118	.217	.319	.356*	.139
Foreign language	.484**	.718**	.211	.201	.260	.316	.495**	.227

Note. ** $p < .01$, * $p < .05$. LI = linguistic intelligence, LMI = logical-mathematical intelligence, SI = spatial intelligence, MI = musical intelligence, KI = bodily-kinesthetic intelligence, INPI = interpersonal intelligence, INTI = intrapersonal intelligence, NI = naturalist intelligence.

Multiple regression analyses were conducted in which the variance caused by the MII was removed, and partial correlations were computed between course grades and children's STAT total and subsection scores. Separate analyses were conducted for each subject area using first the STAT subsections and then using just the STAT total scores. Analyses regarding mathematics course grades yielded significant partial correlations for the creative subsection score ($P_r = .44$, $p < .01$) and for the total STAT score ($P_r = .62$, $p < .01$), but the partial correlations were not significant for the analytical ($P_r = .14$) and practical ($P_r = 05$) STAT scores. Similarly, the regression analyses predicting students' science course grades yielded significant partial correlations for STAT total scores ($P_r = .53$, $p < .01$) and for the creative subsection score ($P_r = .42$, $p < .01$), but not for the analytical ($P_r = .14$) or practical ($P_r = .06$) STAT scores. Additionally, when I performed the same analyses of social science course grades these yielded significant partial correlations with STAT total scores ($P_r = .54$, $p < .01$) and creative subsection scores ($P_r = .34$, $p < .05$) but not with analytical ($P_r = 19$) or creative ($P_r = .04$) STAT scores. Finally, analyses yielded the same pattern for foreign language course grades and STAT total and subsection scores. Regression analyses yielded significant partial correlations for practical subsection scores ($P_r = .41$, $p < .02$) and for total STAT scores ($P_r = .61$. $p < 01$). Thus, the total STAT scores and creative subsection scores significantly predicted academic achievement in mathematics, science, social science, and foreign language courses, independent of multiple intelligences scores; however, the analytical and practical subsection scores did not. Correspondingly, the partial correlations

between course grade (for mathematics, social science, science, and foreign language) and the MII subsection scores, with the variation caused by the STAT removed, were significant only for LMI ($P_r = .70$, $p < .01$) scores. This finding indicates that, independent of the STAT, only LMI scores predicted achievement in any subject area.

Discussion

The results in this study showed that STAT total scores were significantly related to LI, LMI, and INTI scores. Analytical subsection STAT scores were significantly related to LMI scores. Practical STAT subsection scores were significantly correlated only with INTI scores. These results are based on the partial correlations between multiple intelligences and STAT scores. However, I limited the scope of this study to the students' own preferences in regard to their multiple intelligences. In future studies students' intelligence types should be assessed together with the performances of students on related intelligences for different age groups and different subject areas. In the present study mathematics course grades were significantly related to STAT total scores and to scores for the STAT analytical, practical, and creative abilities subsections. Similarly, science, social science, and foreign language course grades were significantly related to the LI, LM, and INTI scores of the participants.

Results of multiple regression analyses indicated that total STAT scores and creative ability scores significantly predicted academic achievement in mathematics, social science, science, and foreign language learning, independent of multiple intelligences scores; however, the analytical and practical ability scores did not. These results are consistent with those reported by Sternberg et al. (2001), who found that total STAT and creative ability scores significantly predicted academic achievement. However, contrary to the findings reported by Sternberg et al., in my study the analytical and practical ability scores did not relate significantly to academic achievement. On the other hand, Koke and Vernon (2003) reported that total STAT scores and only practical ability scores predicted psychology course midterm grades of university students. All these results might indicate that there may be cultural differences within the dominant cognitive abilities represented in the national education systems of various countries.

My results in this study also revealed that the partial correlation between course grades for all of the subject areas and each of the MII subsection scores, with the variation caused by the STAT removed, was significant for only the LMI score. This indicates that, independent of the STAT, only LMI scores predicted achievement in any subject area. It should also be noted that in this study the students' multiple intelligences scores were based on their own preferences for

the items representing various kinds of intelligences. In other words, the multiple intelligences scores did not indicate the actual performance of the children in each type of intelligence. I believe that it would be of value for future researchers to test how well the STAT would predict academic achievement for scores on a test in which students' multiple intelligences scores were each taken into account separately. The relationship between other tests and STAT scores could also be examined with more heterogeneous sample groups.

References

Al-Balhan, E. M. (2006). Multiple intelligence styles in relation to improved academic performance in Kuwaiti middle school reading. *Digest of Middle East Studies, 15*, 18-34. **http://doi.org/cd8zdh**

Armstrong, T. (2009). *Multiple intelligences in the classroom*. Alexandria, VA: ASCD.

Brody, N. (2003). Construct validation of the Sternberg Triarchic Abilities Test: Comment and reanalysis. *Intelligence, 31*, 319-329. **http://doi.org/ffgmzb**

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Castejón, J. L., Gilar, R., & Perez, N. (2008). From "g factor" to multiple intelligences: Theoretical foundations and implications for classroom practice. In E. P. Velliotis (Ed.), *Classroom culture and dynamics* (pp. 101-127). New York: Nova Science.

Douglas, O., Burton, K. S., & Reese-Durham, N. R. (2008). The effects of the multiple intelligence teaching strategy on the academic achievement of eighth grade math students. *Journal of Instructional Psychology, 35*, 182-187.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences*. New York: Basic.

Gottfredson, L. S. (2002). g: Highly general and highly practical. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The general intelligence factor: How general is it?* (pp. 331-380). Mahwah, NJ: Erlbaum.

Greenhawk, J. (1997). Multiple intelligences meet standards. *Educational Leadership, 55*, 62-64.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger/Greenwood.

Koke, L. C., & Vernon, P. A. (2003). The Sternberg Triarchic Abilities Test (STAT) as a measure of academic achievement and general intelligence. *Personality and Individual Differences, 35*, 1803-1807. **http://doi.org/fmfpqb**

McMahon, S. D., Rose, D., & Parks, M. (2004). Multiple intelligences and reading achievement: An examination of the Teele Inventory of Multiple Intelligences. *The Journal of Experimental Education, 73*, 41-52. **http://doi.org/bwptfs**

Mettetal, G., Jordan, C., & Harper, S. (1997). Attitude toward a multiple intelligences curriculum. *Journal of Educational Research, 91*, 115-122. **http://doi.org/dmsgds**

Özdermir, P., Güneysu, S., & Tekkaya, C. (2006). Enhancing learning through multiple intelligences. *Journal of Biological Education, 40*, 74-78. **http://doi.org/fn2x6h**

Reid, C., Romanoff, B., Algozzine, B., & Udall, A. (2000). An evaluation of alternative screening procedures. *Journal for the Education of the Gifted, 23*, 378-396.

Saban, A. (2002). *Öğrenme ve öğretme* [Learning and teaching: New theories and approaches]. Ankara: Nobel.

Sternberg, R. J. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of Personality and Social Psychology, 49*, 607-627. **http://doi.org/cstvmp**

Sternberg, R. J. (1993). *The Sternberg Triarchic Abilities Test*. Unpublished manuscript.

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52*, 1030-1037. **http://doi.org/dzxj2p**

Sternberg, R. J. (1999a). Intelligence as developing expertise. *Contemporary Educational Psychology, 24*, 359-375. **http://doi.org/dzvjsj**

Sternberg, R. J. (1999b). The theory of successful intelligence. *Review of General Psychology, 3*, 292-316. **http://doi.org/cqrkxh**

Sternberg, R. J. (2000). The concept of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 3-13). New York: Cambridge University Press.

Sternberg, R. J. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.

Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities Test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment, 17*, 1-16. **http://doi.org/cn7tjp**

Sternberg, R. J., Ferrari, M., Clinkenbeard, P. R., & Grigorenko, E. L. (1996). Identification, instruction, and assessment of gifted children: A construct validation of a triarchic model. *Gifted Child Quarterly, 40*, 129-137. **http://doi.org/d3rf9w**

Snyder, R. F. (1999). The relationship between learning styles/multiple intelligences and academic achievement of high school students. *High School Journal, 83*, 11-20.

Journal of Physical Education and Sport ® (JPES), Vol 19 (Supplement issue 3), Art 142, pp 987 - 991, 2019
online ISSN: 2247 - 806X; p-ISSN: 2247 – 8051; ISSN - L = 2247 - 8051 © JPES

Original Article

Reliability and validity of the basic motor ability test in preschool children.

CARLOS AYAN[1], SILVIA VARELA[2], MIGUEL A. SANCHEZ-LASTRA[3] AND ÓSCAR MARTINEZ DE QUEL[4]
[1,2,3]Department of Special Didactics, Faculty of Education and Sport Sciences, University of Vigo, SPAIN
[4]Faculty of Education, Complutense Univesity of Madrid, SPAIN

Published online:May 31, 2019
(Accepted for publication April 22, 2019)
DOI:10.7752/jpes.2019.s3142

Abstract:
The Basic Motor Ability Test (BMAT) is a motor ability battery specifically designed for children between 4 and 12 years of age but its psychometric properties have not been thoroughly examined. The purpose of this study was to analyze the feasibility, reliability and validity of the BMAT when administered to preschool children. Spanish healthy children (N = 75) from three different kindergarten schools were assessed on two occasions separated by four weeks. Five BMAT subtests showed high test-retest correlations (0.75-0.86) whereas the rest revealed a weak to moderate reliability (0.48-0.64). BMAT internal consistency was found to be weak (Cronbach's Alpha = 0.49). All the subtests correlated with age. The BMAT is a feasible motor assessment tool that can be performed by preschoolers, albeit with some modifications. The lack of reliability reported in several subtests is an important concern that should be confirmed by future studies with larger samples.
Key Words: - Psychometric properties, field-based battery, motor skills, child motor development, human performance

Introduction

The assessment of motor ability in young children has become increasingly important in recent years, since it has been suggested that it could be linked to cognitive, language, social and emotional difficulties (Piek, Hands, & Licari, 2012). Moreover, there is evidence that poor motor ability may impact on physical fitness (Barnett, Van Beurden, Morgan, Brooks, & Beard, 2008), an important marker for health and disease during childhood (Ortega et al., 2014). Therefore, there is a need for accurately measuring motor ability during the preschool years.

Motor ability in preschoolers is often assessed by means of norm-referenced field-based batteries that should meet some criteria such as easy and fast administration, low cost, appropriate psychometric properties and simplicity in the calculation of a final score (Cools, De Martelaer, Samaey, &Andries., 2009). In this respect, the Basic Motor Ability Test (BMAT) (Arnheim & Sinclair, 1975) is a motor ability battery specifically designed for children between 4 and 12 years of age. Although BMAT accomplishes most of the requirements mentioned earlier, there are three important issues that need to be considered regarding its use. First, two of its nine subtests, Bead Stringing and Tapping, require specific instruments to be implemented. Although they can be adapted by using other non-specific materials, it is unknown how this would affect their reliability. Secondly, as the normative values for the BMAT were established 40 years ago, its current validity is unclear since preschoolers' motor development might have changed during this period (Runhaar et al., 2010). Furthermore, there might be differences related to the sociodemographic characteristics of the population studied (Kambas et al., 2012). Lastly, while the overall psychometric properties of the BMAT have been reported (Arnheim & Sinclair, 1975), its reliability and validity for different age groups remain unestablished. This is of special relevance because preschoolers may have difficulties in understanding the test protocol or may lack in motivation and this could affect the psychometric properties of the BMAT (Ayán, Cancela, Romero, & Alonso, 2015).

Given these issues, the present study aims to identify the feasibility, reliability and validity of the BMAT when administered to preschool children.

Material & methods

Participants

The participants were Spanish healthy children from three different kindergarten schools. Children enrolled in the second level (4 and 5 years old) of the first period of the Spanish Education Curriculum were

study, formal permission from the principals of the schools involved was sought and granted. Written informed consent was obtained from the parents/guardians of all children and assent was obtained from the minors previously to their participation in the study. The regional ethics committee reviewed and approved the protocol of the study.

Measures

Anthropometry. Body weight was measured to the nearest 0.1 kg using a digital scale (Tefal PP1200VO) and height was measured to the nearest millimetre with a field stadiometer (Seca 220). Weight and height were measured in light clothing without shoes. Body mass index (BMI) was calculated as weight in kilograms divided by height in meters squared (kg/m2).

Basic motor ability test (BMAT). This battery is composed of the following nine subtests:

Bead stringing (BS). This subtest consisted in threading beads (1 cm in diameter) onto a cord (45 cm long with a plastic end of 2 cm in diameter) as rapidly as possible during 30 s. Given that the Stanford-Binet test beads recommended by the standard protocol of the BMAT weren't available, screws with identical dimensions were used.

Target throwing (TT). In this subtest, children were required to throw 15 squared bags of seeds (10-12.5 cm wide) towards three targets (placed on a wall at a height of 1.2 m), from a distance of 2 m. The targets consist of three rectangles with heights of 12, 27 and 45 cm. Each target is given a different score according to its size so that 1, 2 and 3 points are awarded to the large, medium and small rectangles, respectively. Once the 15 bags have been thrown, the final score was summed up for each participant.

Tapping (T). This subtest consisted in hitting alternatively two circles located on an electronic board (45 cm long), during 20 seconds. Given that the specific electronic board wasn't available and following the BMAT guidelines, a 45 cm long board was built of cardboard and the two circles were drawn on the corners. The examiner was responsible for keeping track of hits.

Hamstring stretch (HS). In this subtest, the children were required to sit on the floor with both legs outstretched and heels 15 cm apart from each other. A 3 m ruler or stick should have been placed in the middle of both legs with the 30 cm mark aligned with the heels, but as there was none available, a measurement tape was fixed to the ground at the corresponding position. Without bending their knees, children should try to reach forward as far as possible with their fingertips. The distance reached was recorded in centimeters.

Long jump (LJ). In this subtest, the children had to perform 3 standing long jumps on a level and non-slip floor surface. A take-off line was marked with tape on the floor. At the start of the jump the feet must be behind the take-off line. There must be no movement of the feet prior to take-off and the children must retain balance after landing. Jump distance was measured as the distance from the take-off line to the nearest body part upon landing (this is typically the point of heel contact). The best of the three trials was recorded.

Face down to standing (FD). This subtest began with the child lying upside down on a mat (1.2 x 1.8 m). At the examiner's signal "Ready? Go!" the child had to stand up and go back to the initial position as many times as possible during 25 s. The number of times the child was able to complete this whole action was registered.

Static balance (SB). In this subtest the children were asked to stand on one foot as long as possible on a 2.5 cm wide wooden platform (SB1). Children were instructed to close their eyes, to place their hands at the waist and to hook the lifted foot behind the knee of the supporting leg. The test finished when the participant sat the lifted foot on the floor, opened the eyes or removed the hands from the waist. A 10 s trial repetition was allowed before the test was recorded. According to the BMAT guidelines, this subtest was repeated on a 5 cm wide wooden platform (SB2).

Push-ups (PU). In this subtest, children had to do as many arm push-ups as they could, setting both feet on the floor and placing both hands on a bench so the body was in inclined position.

Agility run (AR). This subtest consisted in running in a zig-zag pattern between four cones spread 1.5 m apart in a straight line. The starting point was at the right side of the first cone. Children had to perform as many zig-zag runs as they could in 20 seconds. The number of runs achieved represented the final score.

Procedure

All subtests were carried out in groups of 20 children on a four-week schedule. In the first week, BMI was determined and the BMAT protocol was carefully explained to the children to avoid learning effects during the experiment. In the second week, the participants performed the BMAT (test). During the third week no assessments were done. Finally, BMAT was carried out again in the fourth week (retest). Three senior students who were majoring in early childhood education administered the tests and a kindergarten teacher supervised the assessments. They were all specialized in Physical Education.

Statistical Analysis

Data were analyzed in several stages. First, assumptions of normality and homoscedasticity were checked for each dependent variable using a one-sample Kolmogorov-Smirnov test and Levene's test, respectively. Secondly, descriptive statistics were calculated and a comparison of means was carried out using an independent samples Student's t-test to examine the differences between sexes. Thirdly, reliability was assessed for each BMAT subtest by calculating the test-retest correlation. Pearson's and Spearman's correlations were used for

Cronbach's alpha. Lastly, construct validity was established by correlating BMAT subtests scores with age since it was expected that preschoolers' motor skills improve with maturation. SPSS 15.0 for Windows was used for statistical analysis (SPSS Inc., Chicago, IL, USA) and statistical significance was established at $p<0.05$.

Results

A total of 75 children (mean age 5.51 ± 0.29 years) volunteered to participate and completed the study. Participant characteristics are summarized in Table 1 along with their performance in the different BMAT subtests. Significant intersexual differences were observed only for flexibility, both in the test and the retest ($p<0.05$), with girls presenting higher values of HS than boys.

The battery showed to be feasible as children found no difficulties in understanding the tasks that were to be executed. Regarding task execution, children found difficult to perform the SB and the PU subtests since it took some time to achieve their correct realization pattern.

Concerning the reliability of the BMAT, five subtests (BS, T, HS, LJ and PU) showed high test-retest correlations (0.75-0.86), whereas the rest of the subtests revealed a weak to moderate reliability (0.48-0.64) (Table 2).

In relation to the internal consistency of the battery, the Cronbach's Alpha coefficient was 0.49, thus suggesting a weak correlation among the different subtests that the BMAT comprises.

Finally, regarding construct validity, all the subtests directly correlated with age, except HS, which was inversely correlated, as expected (Table3). The TT was the only subtest that showed a significant statistical association with age.

Dicussion

Reliability and validity are necessary features of a good instrument for the assessment of motor ability. So this study tried to identify if the BMAT proved to be valid and reliable when administered to preschoolers. The obtained results may be used to assist any physical education teacher or health professional in establishing children´s level of motor development or in ascertaining the effects of interventions on children's motor ability.

The BMAT proved to be feasible except for the balance subtest, which showed a high level of difficulty reflected by the short time that the children were able to hold the required body posture, regardless of platform width. It also needs to be pointed out how difficult it was for the children to correctly execute the PU subtest, a common issue in previous studies (Ayán, Cancela, Senra, & Quireza, 2014). The feasibility of the BMAT when administered in preschool settings was ratified by the fact that the adjustments made to the materials used in the BS and the TT subtests did not affect its execution.

The present study also allowed us to compare the current sample mean scores with normative data from the original manual of the BMAT published 40 years ago. Both boys and girls were on the 50th percentile in BS, TT, LJ and FD subtests and both surpassed the 90th percentile in T and AR. In HS and PU the children scored close to the 25th percentile but none of the participants reached this percentile in SB. Given that the original manual of the BMAT did not provide detailed information about the sample employed to establish the normative curves, the differences observed in some subtests are acceptable. Only the scores in SB can be considered an unusually low level of performance. The reason that would explain this extremely low balance scores is the difficulty of the subtest because standing on one foot with eyes closed on a narrow platform can be excessively demanding for preschoolers.

The difficulty of the task involved in the SB subtest could have also affected its reliability, given that similar balance tests have demonstrated to be easier for preschoolers and have shown acceptable reliability when executed on the floor (Crock, Horvat, & McCarthy, 2001; Fjortof, 2010; Larkin & Revie, 1994; Mc Carron, 1997), or on a platform (Bös, Bappert, Tittlbach, & Wall, 2004; Klein, Koch, Dordel, Strüder, & Graf, 2012), with eyes open. In this line, the low reliability observed in some other subtests of the BMAT (AR, TT and FD) could be attributed to their protocol characteristics. For example, the agility subtest comprises several changes in direction in a 6 m distance and it has been noted that including many changes in direction and a long distance are questionable aspects of an agility test (Sayers, 2015). In fact, all of the agility tests that have proved to be reliable for children less than six years of age propose a shorter distance (4m) and fewer changes in direction (Ortega et al., 2014). Analogously, the TT subtest asked the children to throw toward a concentric rectangular target, while it has been observed that there is a higher reliability in tests with circular targets (Malina, 1968). Moreover, the reliability on this kind of tests strongly depends on the target size and on the throwing distance (Zahradník, Vaverka, & Gajda, 2008). In addition, the BMAT protocol does not consider the participant physical capacities and anthropometric characteristics, as the dimensions and distances established for the TT subtest does not change with age. It could be expected that the TT subtest would be more reliable if this requirement were fulfilled and target dimensions were bigger for the younger children, as in other motor development assessment batteries (Bruininks, 2005; Zimmer & Volkamer, 1987). Finally, concerning the FD subtest reliability, the BMAT guidelines state that the face down to standing task demands speed and agility from the participants, but after analyzing their execution it became clear that strength and resistance are also necessary capacities in order to perform the subtest correctly. As a consequence of the great effort required to complete the FD, motivation

reliability can be affected if the task is perceived to be exhausting or when it is not known how to meet its physical or conditional demands (Ayán et al., 2014).

On the other hand, The BAMT comprises two subtests (LJ and BS) that showed a good reliability. The LJ is a reliable test commonly used to assess lower body muscular strength in young children (Ortega et al., 2014). Similarly, despite the adjustments made in this study, the BS also resulted to be reliable, as previously observed (Crock et al, 2001). In this line, the T subtest showed an acceptable reliability after the modifications implemented in this research. With respect to this finding, it is worth mentioning that Fjortof (2010) suggested that the protocol of the Tapping test included in the EUROFIT battery, which is similar to the T subtest in the BMAT, should be modified due to the its poor reliability when applied to young children. In addition, it is important to point out the high reliability observed for both PU and HS subtests, two of the very few field-based tests that evaluate upper limb strength and flexibility without the need of special equipment, since to the authors' knowledge it has not been previously reported in preschoolers.

The BMAT battery has been regarded as a valid tool by its creators and given the positive correlations found in the present study between subtests scores and participants' age, the BMAT demonstrated to have acceptable criterion validity. In this regard, it is important to highlight that the correlation was only significant for the flexibility subtest and this could be due to the fact that this physical capacity worsens rapidly with age. The narrow age range of the sample (60-72 months) could explain the absence of other statistically significant associations. Indeed, a greater number of statistically significant associations would be expected simply by covering a broader age range.

The BMAT showed a weak internal consistency because each subtest measures different physical capacities (e.g. strength, agility, balance, etc.). In contrast, a high level of internal consistency (Cronbach's alpha = 0.89) has been reported in a previous study (Kavianpour, Raki, & Malekpour, 2014), but its sample consisted of only three preschoolers who were diagnosed with developmental coordination disorder.

Several studies have reviewed the psychometric properties of various batteries assessing motor ability (Cools et al., 2009; Slater, Hillier, & Civetta, 2010; Piek et al., 2012; Wiart & Darrah, 2001), but none of them have thoroughly analyzed the BMAT. The results provided here represent an initial attempt to remedy this lack of information. However, the small sample size and the fact that no data regarding the criterion validity of the battery have been incorporated are two important limitations that must be considered when interpreting them.

Conclusions

This investigation shows that the BMAT can, with certain modifications, be applied to 4-6 year old children, but its effectiveness in evaluating young children's motor development is in doubt. Professionals who are willing to apply the BMAT should take into account that some items, especially the balance subtest, can be difficult to perform by preschoolers, while others, such as the flexibility or the tapping subtests, seemed to be very reliable and easier to carry out. Consequently, the most appropriate option for physical education and psychomotor activity professionals would be to apply this later BMAT subtests and to use other alternatives to evaluate those mobility components in which the BMAT showed a weak reliability.

Conflicts of interest - The authors report no conflicts of interest.

References:

Arnheim, D.D., & Sinclair, W.A. (1975). *The clumsy child: A program of motor therapy*. England: C. V. Mosby.

Ayán, C., Cancela, J.M., Romero, S., & Alonso, S. (2015). Reliability of two field-based tests for measuring cardiorespiratory fitness in pre-school children. *The Journal of Strength & Conditioning Research, 29*(10), 2874-2880.

Ayán, C., Cancela, J.M., Senra, I., & Quireza, E. (2014). Validity and reliability of 2 upper-body strength tests for preschool children. *The Journal of Strength & Conditioning Research, 28*(11), 3224-3233.

Barnett, L., Van Beurden, E., Morgan, P., Brooks, L.O., & Beard, J.R. (2008). Does childhood motor skill proficiency predict adolescent fitness? *Medicine & Science in Sports & Exercise, 40*(12), 2137-2144.

Bös, K., Bappert, S., Tittlbach, S., Woll, A. (2004). Karlsruher Motorik-Screening für Kindergartenkinder (KMS 3–6) [Karlsruher motor screening for kindergarten children (KMS 3-6)]. *Sportunterricht, 53*, 79–87.

Bruininks, R.H. (2005). *Bruininks-Oseretsky Test of Motor Proficiency, (BOT-2)*. Minneapolis, MN: Pearson Assessment.

Cools, W., De Martelaer, K., Samaey, C., & Andries, C. (2009). Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. *Journal of Sports Science & Medicine, 8*(2), 154-168.

Crock, R.V., Horvat, M., & McCarthy, E. (2001). Reliability and concurrent validity of the movement assessment battery for children. *Perceptual & Motor Skills, 93*(1), 275-280.

Fjortoft, I. (2010). Motor fitness in pre-primary school children: The EUROFIT motor fitness test explored on 5-7-year-old children. *Pediatric Exercise Science, 12*(4), 424-436.

Kambas, A., Venetsanou, F., Giannakidou, D., Fatouros, I.G., Avioniti, A., Chatzinikolaou, A., ...Zimmer, R. (2012). The Motor-Proficiency-Test for children between 4 and 6 years of age (MOT 4–6): An investigation of its suitability in Greece. *Research in Developmental Disabilities, 33*(5), 1626-1632.

Kavianpour, F., Raki, A., & Malekpourm M. (2014). Efficacy of training of executive functions (working memory) on the rate of attention in preschool children with developmental coordination disorder. *Zahedan Journal of Research in Medical Sciences, 16*(9), 89-94.

Klein, D., Koch, B., Dordel, S., Strüder, H., & Graf, C. (2012). The KiMo-test: a motor screening for pre-school children aged 3–6 years. *Gazzetta Medica Italiana, 171*(1), 13-26.

Larkin, D., & Revie, G. (1994). *Stay in step: A gross motor screening test for children K-2*. Perth, Australia.

Malina, R.M. (1968). Reliability of different methods of scoring throwing accuracy. *Research Quarterly, 39*(1), 149-160.

McCarron, L.T. (1997). *MAND McCarron Assessment of Neuromuscular Development: Fine and gross motor abilities*. Dallas, TX: Common Market Press.

Ortega, F.B., Cadenas-Sánchez, C., Sánchez-Delgado, G., Mora-González, J., Martínez-Tellez, B., Artero, E.G., ...Ruiz, J.R. (2014). Systematic review and proposal of a field-based physical fitness-test battery in preschool children: The PREFIT battery. *Sports Medicine, 45*(4), 533-555.

Piek, J.P., Hands, B., & Licari, M.K. (2012). Assessment of motor functioning in the preschool period. *Neuropsychology Review, 22*(4), 402-413.

Runhaar, J., Collard, D.C.M., Singh, A.S., Kemper, H.C., van Mechelen, W., & Chinapaw, M. (2010). Motor fitness in Dutch youth: differences over a 26-year period (1980–2006). *Journal of Science & Medicine in Sport, 13*(3), 323-328.

Sayers, M.G. (2015). The influence of test distance on change of direction speed test results. *The Journal of Strength & Conditioning Research, 29*(9), 2412-2416.

Slater, L.M., Hillier, S.L. & Civetta, L.R. (2010). The clinimetric properties of performance-based gross motor tests used for children with developmental coordination disorder: a systematic review. *Pediatric Physical Therapy, 22*(2), 170-179.

Wiart, L., & Darrah, J. (2001). Review of four tests of gross motor development. *Developmental Medicine & Child Neurology, 43*(4), 279-285.

Zahradník, D., Vaverka,, F., & Gajda, V. (2008). Optimisation of the size of a target and the throwing distance during a throw at a target for adults. *Universitatis Palackianae Olomucensis Gymnica, 38*(4), 39-45.

Zimmer, R., & Volkamer, M. (1987). *Motoriktest für vier-bis-sechsjährige Kinder* [Motor test for four to six years old children]. Betz, Ukraine: Weinheim.

Psychoanalytic Psychology
2019, Vol. 36, No. 1, 44–52

0736-9735/19/$12.00 http://dx.doi.org/10.1037/pap0000167

Bringing a Psychoanalytic Mindset to Neuropsychological Testing: From Parameters and Testing the Limits to the "Something More"

Sharon Leak, PhD
Pittsburgh Psychoanalytic Center, Pittsburgh, Pennsylvania

If one understands the projective hypothesis most simply as "the active structuring of the world according to inner requirements and outer demands" (Schafer, 1954, p. 37), then it is evident that a patient's response to an ostensibly neuropsychological measure can provide significant insights beyond the typical neurocognitive yield. Indeed, the more ambiguous the instructions or stimuli, the greater the potential for projective elements to be introduced. The Wisconsin Card Sorting Test offers such an opportunity, particularly in its more recent computerized version. In this paper, a clinical vignette illuminates how bringing a psychoanalytic mindset to the administration and interpretation of this widely used measure can enrich its diagnostic utility. A distinction is made between a parameter (as applied to psychological testing) and testing the limits, with appreciation for how both can lead to vivid clinical moments that inform diagnostic questions—a "something more." Such interventions maintain the essence of a standardized administration without allowing technical rigidity to obscure deeper contact with, and understanding of, the patient.

Keywords: Wisconsin Card Sorting Test, parameter, testing the limits, "something more"

Given a history of personality assessment that is roughly contemporaneous with Freud's early psychoanalytic writings in 1896, as well as assessment's focus on disentangling and integrating threads of data from intelligence, achievement, neuropsychological, and personality measures, it becomes clear that psychodynamic principles must play a central role in any testing that purports to look at the patient, first and foremost, as a person (Bornstein, 2010; Bram & Yalof, 2015; Leak & Hayden, 2015; Meersand, 2011). And yet frequently there is an atheoretical approach to testing that compromises all aspects of the psychological assessment process, including selection of test measures, administration procedures, interpretation of data, and feedback.

Neuropsychological testing is especially vulnerable to being divorced from psychodynamic principles. Not only is there the declining influence of psychoanalytic thinking in the typical psychology curriculum and the pejorative tone that may accompany what is presented (Bornstein, 2001), but also one finds the "pervasive problem in contemporary psychology" of relying heavily on self-report measures (Bornstein, 2001, p. 135). Thus, psychologists who carry out neuropsychological testing can become disproportionately familiar with actuarial and descriptive approaches.

When contemporary trends in education and practice are combined with a lack of exposure to, or awareness of, key functions served by psychoanalytic theory in diagnostic psychological testing, the result is a greatly reduced ability to make use of the multiple rich sources of information generated through the assessment process. As Sugarman and Kanner (2000) point out, psychoanalytic theory serves organizing, integrating, clarifying, and predictive functions by "recasting psychological test data as psychological constructs whose relationship is already delineated by the psychoanalytic theory of the mind" (p. 6).

Without this theory of mind and a grasp of internalized object relations, how does one understand and make use of transference–countertransference responses that emerge in the course of a psychological assessment? Without an appreciation of the interplay between conscious and unconscious processes, how does the evaluator understand instances where patients' self-report data converge *or* contrast with data from neuropsychological or projective measures? Without being able to create a testing alliance that recognizes the value of allowing patients "to usefully flounder a bit," enabling them to become aware of their own resilience while also clarifying the nuances of the evaluator's diagnosis (Tuber, 2012, p. 229), how can the psychologist achieve the goal of arriving at a valid and useful diagnostic conceptualization?

Questions such as these prompted the writing of this paper. My aim is twofold: First, I would like to contribute to an ongoing dialogue regarding the value of psychoanalytic theory in psychological testing, especially with respect to neuropsychological evaluations, where psychodynamic principles often are viewed as irrelevant. Second, my desire is to provide a nuanced glimpse into the rich potential of this assessment process as it is applied to a patient in psychoanalytic psychotherapy.

This article was published Online First November 27, 2017.

Sharon Leak, PhD, Pittsburgh Psychoanalytic Center, Pittsburgh, Pennsylvania.

I thank Gerrian Bobrowsky, PhD, who taught the Writing Class at Pittsburgh Psychoanalytic Center; and my fellow candidates, for comments on an initial draft of this paper.

Correspondence concerning this article should be addressed to Sharon Leak, PhD, 2345 Murray Avenue, Suite 230, Pittsburgh, PA 15217. E-mail: drsaleak@gmail.com

Converging Lines of Thought From the Literature

Early Neuropsychological Assessment

Psychological assessment of a predominantly neuropsychological nature was not under the purview of psychoanalysts when the tradition of Rapaport, Gill, and Schafer (1968) came to the fore, so it is understandable that even an early, classic edition of *Neuropsychological Assessment* (Lezak, 1983) fails to explicitly mention the need for psychodynamic principles to guide such testing. And yet the approach the author outlines remains pertinent, with its appreciation for the complexity of diagnostic psychological testing and a sensitive attunement to the patient that is in keeping with the "entwined history" of personality assessment and psychoanalysis (Bram & Yalof, 2015, p. 1):

> Standardized procedures eliciting behavior that can be measured along empirically and scaled dimensions provide objectivity and the potential to make fine distinctions and comparisons which would be unattainable by clinical observation alone. Still, examinations cannot be adequately conducted nor can test scores be properly interpreted in a psychological or social vacuum. The uniqueness of each patient's capacity, disability, needs, and situation calls for discriminating, flexible, and imaginative use of examination techniques (p. 4).

Personality Assessment in the Tradition of Rapaport, Gill, and Schafer

A psychoanalytically informed tradition of testing was explicitly captured in Rapaport et al.'s 1968 classic *Diagnostic Psychological Testing*. And though today's assessment battery has been expanded and updated to reflect evolving theory and research, the informed practitioner continues to select measures assessing intelligence, achievement, neuropsychological functioning, and personality that best address the referral question at hand (Bram & Peebles, 2014; Rothstein, Benjamin, Crosby, & Eisenstadt, 1988; Rothstein & Glenn, 1999; Tuber, 2012; Yalof, 2006).

Today's test battery need not eschew self-report measures, but such results must be integrated with findings from neurocognitive and personality measures to gain insight into conscious and unconscious processes. Further, all measures are scored, analyzed, and interpreted in the context of the relational process that unfolds between the evaluator and patient. As Bram and Yalof (2015) point out, personality assessment complements the idiographic approach of psychoanalysis by integrating it with a nomothetic approach that applies "quantitative methods to determine in what ways and to what extent a person is similar or different relative to normative data" (p. 75). Giving equal weight to qualitative and quantitative data frees the evaluator to refine diagnostic hypotheses through sensitive inquiry on test items, judicious alteration of procedures, testing-the-limits strategies, or monitoring of the countertransference. Quite simply, all data are considered through a psychoanalytic lens—a lens that is ever evolving.

A classic paper by Joel Allison (1978) illuminates how neurocognitive findings are enriched by a psychoanalytic mindset. In "Clinical Contributions of the Wechsler Adult Intelligence Scale," he scrutinizes patient responses to intelligence testing within an ego psychological framework. Although this Wechsler is now in its fourth edition (2008), Allison's historical grounding and astute clinical analyses remain pertinent. The development of ego psychology, he commented, allowed for an extension of the earlier projective hypothesis: Psychological assessment could continue to focus on an individual's unconscious wishes, fantasies, conflicts, and motivations while also exploring "aspects of style that were organized into particular patterns" (p. 355).

More recent efforts in the testing field have continued to articulate the manner in which evolving psychoanalytic models can inform and integrate neuropsychological and personality assessment (Bornstein, 2010; Bram & Peebles, 2014; Bram & Yalof, 2015; Rothstein et al., 1988; Rothstein & Glenn, 1999; Tuber, 2012; Yalof, 2006). This broad range of psychoanalytic thinking can be used to glean nuanced data from the content and process of a psychological assessment. Together with attempts to reconnect psychoanalysis with mainstream psychology (Bornstein, 2005), evolving theory and technique should enrich the psychodiagnostic testing enterprise rather than reduce it to easily scored self-report measures of limited utility.

Parameters, Testing-of-Limits Strategies and the "Something More" in Testing

It is not only that clinical data can be best heard and understood through a psychoanalytic lens, but also that test administration procedures must be flexibly carried out for optimal assessment. Here, the concept of a parameter—Eissler's (1953) guide to deviations in the analyst's (then) model technique of interpretation as the exclusive tool in psychoanalysis—provides an analog[1] for test administration. This classic paper notes four conditions for such a deviation, two of which are pertinent here: First, the parameter should be introduced only when the model technique is not sufficient; second, the parameter should not exceed the unavoidable minimum. Although clear differences exist between psychodiagnostic testing and psychoanalysis, there are also parallels regarding how one can flexibly meet a patient's needs without compromising the essential task and boundaries of the clinical encounter.

Definitions

Let me first articulate a subtle but critical distinction between testing the limits and introducing a parameter within a psychological evaluation. As Bram and Peebles (2014) point out, standardized administration is essential to reliable scoring. When testing the limits, it is only after a scorable response has been obtained according to standardized procedures that the evaluator is free to test the limits through "incremental assists"—graded interventions designed to discover what helps a patient "refocus, regroup, restabilize, or problem solve" (p. 70). These gradually increasing levels of support or direction yield alternate (not standardized) scores.

What happens, however, if a test is not amenable to incremental assists due to its length or the potential for assists to intrude on standardized scoring? Such assists are invaluable on Wechsler subtests, where, for example, it is possible to score as incorrect a response not completed within the time limit while allowing the patient to continue a bit longer to see if extended time allows for correct completion. Also, I routinely circle operational signs and ask patients to redo problems on math achievement tests if errors

[1] I am grateful to Erwin Flaxman, PhD, for bringing this analog to my attention.

reflect use of a different sign: I want a clear assessment of calculation abilities apart from perceptual or attentional challenges; I then compute standardized and alternate scores.

Contrast these examples with the Conners' Continuous Performance Test II (CPT-II, Conners, 2002), a 15-min computerized test that assesses attention; the patient is asked to click the mouse each time a letter appears on the screen, except if the letter is *X*. Here, standardized administration allows for a single reorienting prompt for any deviation, including the patient leaving his or her seat. I recall a moment from the supervision of a young psychologist who was puzzling with me over how to interpret the results of a boy who had jumped up midtest, announcing he had to go to the bathroom. He ran down the hall to the men's room and returned several minutes later to complete the test, which had continued to generate letters in his absence. The trainee explained: "I had already given him the single prompt to stay in his seat and I wanted this to be a valid administration."

Not surprisingly, the boy's profile and discriminant function analysis suggested profound attention problems, but these formal scores were meaningless and the failure to systematically intervene based on evaluator hypotheses left us with no data from a graded series of interventions. Here, the nature of the test does not allow for both standardized and alternate scores; even if time had been available for a second administration, this would compromise validity. It is most problematic when an assist essentially replicates the function one is measuring—that is, when the evaluator orients a patient's attention to a test designed to assess attention or asks a patient to reflect on the response process when self-monitoring aspects of executive functions are being evaluated. Altering standardized procedure while knowing that scores can only be interpreted in the context of the nonstandardized administration is, in my mind, a parameter—distinct from a testing-the-limits strategy. The evaluator introduces the intervention only when the standardized procedure is not sufficient, and interventions ideally do not exceed the unavoidable minimum.

What testing-the-limits strategies and parameters *do* share is a potential to create during testing a "something more" that involves "special 'moments' of authentic person-to-person connection" that constitute "implicit relational knowing" (Stern et al., 1998, p. 904). In contrast to knowledge represented in verbal or imagistic form, this knowledge captures a way of being with someone that offers new experience and understanding, perhaps analogous to insights derived from subtle enactments in psychoanalytic treatment.

In summary, an evaluator must have a comfortable mastery of standardized administration procedures and remain cognizant of how departure from a protocol can compromise interpretation of results relative to normative data. But rigid adherence to procedures can obscure more than illuminate. A judicious parameter may produce the richest yield of qualitative and quantitative material. After introducing the Wisconsin, I will offer such a moment in a patient's psychological assessment.

The Wisconsin Card Sorting Test (WCST)

The Wisconsin Card Sorting Test (Heaton, Chelune, Talley, Kay, & Curtiss, 1993) was developed to assess abstract reasoning capabilities and the ability to shift cognitive strategies in response to changing environmental contingencies (Grant & Berg, 1948). Standardized for use with children, adolescents, and adults, this test has become recognized during its almost 60 years of use as a multifaceted measure of executive functioning: The Wisconsin requires strategic planning and organized searching alongside effective self-monitoring, including the ability to use feedback to shift cognitive set and inhibit impulsive responding (Heaton et al., 1993). The test has generated great interest (Strauss, Sherman, & Spreen, 2006) because it moves beyond task success or failure to provide data on multiple aspects of problem solving.

Test Materials

The Wisconsin is now available in two modalities: The original version uses four stimulus cards ("key" cards) and decks of response cards, with the evaluator sitting across the table from the patient while providing verbal feedback regarding right or wrong answers; the computerized version presents the key cards horizontally across the top of the computer screen, with the response cards emerging one at a time at the bottom. Following each matching attempt, a "banner" indicating RIGHT or WRONG appears on the computer screen, accompanied by a male voice providing auditory feedback; the evaluator sits to the side and slightly behind the patient. The pace of responding is controlled by the examinee. The four key cards display one red triangle, two green stars, three yellow crosses, and four blue circles. (See Figure 1).[2]

Test Administration

Instructions from the original version have been reworded for consistency with the computerized format (Heaton et al., 1993, p. 5). The evaluator explains:

> This test is a little unusual because I am not allowed to tell you very much about how to do it. You will be asked to match each card that appears at the bottom of the screen to one of these four key cards. Using this mouse you must click on the key card you think it matches. I cannot tell you how to match the cards, but the computer will tell you each time whether you are right or wrong. If you are wrong, simply move on to the next card and try to get the next card correct. If, however, you have clicked on the key card and you change your mind before your card has settled into place, you may click in this outside space and your card will return to the bottom of the screen and you can change your answer. There is no time limit on this test. Are you ready? Let's begin.

The first correct sorting category is Color. Each time the patient responds, the computer generates the banner and auditory response of RIGHT for a correct choice and WRONG for an incorrect choice. This process continues until the patient has produced 10 consecutive Color responses. Without prior indication, the computer changes the correct sorting category to Form (shape), which now remains the correct sorting category until 10 consecutive correct responses are attained. Again without prior indication, the computer changes the correct sorting category, this time to Number. After 10 consecutive correct responses, there is another shift, with Color, Form, and Number presented a second time. Through-

[2] Reproduced by special permission of the publisher, Psychological Assessment Resources, Inc. (PAR), 16204 North Florida Avenue, Lutz, Florida 33549, from the Wisconsin Card Sorting Test by David A. Grant, PhD, and Esta A. Berg, PhD, Copyright, 1981, 1993, by PAR. Further reproduction is prohibited without permission of PAR.

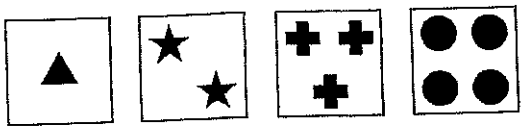

Figure 1. WCST stimulus ("key") cards.

out the test, a stimulus card can match the key card on one, two, or three of these dimensions but there remains only one correct sorting category.

According to standardized administration procedures, at no time should the evaluator provide the patient with any information that is not contained in the initial instructions (Heaton et al., 1993). Patients vary in the amount of time they take to complete the test, but most are expected to do so within 20 to 30 min, either successfully working their way through all six categories or reaching the limits of the 128-card administration. The computerized version of the Wisconsin then provides a printout of individual patient responses and a set of standardized scores (presented with patient scores as part of the clinical vignette in the latter half of this paper).

A Psychoanalytic Mindset

After giving the instructions, the evaluator remains quietly present, allowing both individuals to be in touch with the patient's characteristic mode of responding. Recall of test directives, how the mouse is handled, responses to success and failure: I note these and other qualitative aspects of a patient's response process on a supplementary form[3] devised for this purpose. I observe, for example, whether the patient recalls the "pull-back" option that allows for a change of response if an error is recognized before the card settles into place. The click of the mouse may be timid or aggressive. Or a patient may inquire about the "voice" that the computer ensures is constant in tone and volume: "Why is the voice that says WRONG so much louder than the one that says RIGHT?"

Throughout the test, I silently try to clarify whether there are response tendencies that I need to address to optimally understand a patient's functioning, or which call for a parameter. The computerized administration includes several features that warrant special attention. First, the program will not accept a click of the mouse for the next response until the RIGHT/WRONG banner has receded. Some patients respond so slowly that they never discover this fact. I wonder, does this reflect characterological deliberateness or neurocognitive slowing? Other patients respond prematurely once or twice and then recognize the need to wait for the banner to recede. Still others impulsively click away, either impatient with (or oblivious to) the time lag required for the response to be accepted. After five premature responses, I make this feature explicit to the patient; I then observe how this new knowledge affects subsequent performance.

Similarly, if the patient is not matching to the top row of key cards, I provide redirection on par with the original version of the test to ensure that the patient obtains correct feedback. I might also ask a patient to pause if he or she appears paralyzed or agitated, saying, "Tell me, how are you making your match?" Or I might say to a bright patient verbalizing complex matching schemes, "Keep it simple." In each case, I monitor the patient's response to my intervention to determine the impact of the parameter.

After the test has been completed, I carry out inquiry to assess the patient's capacity for self-reflection, gathering qualitative information about the executive functions the Wisconsin is designed to assess. I routinely ask: "What did you think was happening? When did you change how you were making your match? Did you notice that you had to wait for the banner to recede (before I pointed it out)? Is there anything else you noticed?" I do not, however, provide feedback about these perceptions.

A confidentialized vignette illustrates a deviation from standardized procedure that was introduced during the Wisconsin, allowing for an illuminating clinical moment. A brief overview of the patient's background provides a context for the conditions under which a parameter led to an enriching "something more."

Clinical Vignette: "Mr. A"

Referring Concerns

Mr. A, a 31-year-old single man, was referred for psychological evaluation by an experienced male psychologist, a psychodynamically trained clinician with a sensitivity to learning and attention problems. Mr. A had entered psychotherapy to gain insight into troubling sexual preoccupations that stood in the way of his wish to marry and have children. At the time of the referral, he had begun to experience problems with attention and memory that intruded on his work and social life. Recalling that his father and brother had similarly struggled, and finding himself in the "probable" range of an online screening for attention-deficit disorder (ADHD), he wondered if stimulant medication might help. Following discussion of this "self-diagnosis" with his therapist, the dyad decided to seek formal assessment to clarify whether ADHD, trauma, conflict, or a combination of factors contributed to Mr. A's experience of "not remembering."

Early Childhood

Mr. A reported that he was the youngest of four children in a "suffocatingly close" family. His parents and numerous extended family members worked for the same organization, and it was expected that the patient and his brother (but not his sisters) would eventually join the agency. Growing up, his family took trips together, vacationed together, and spent free time visiting relatives. Mr. A felt that to want friends, a career, or a family of his own would constitute a betrayal.

Mr. A's development was notable for largely typical attainment of early milestones. A sensitive and preoccupied child who was neither impulsive nor overly active, he had difficulty separating and making transitions. Although this early history was not consistent with a "classical" picture of ADHD, it allowed for questions about a predominantly inattentive presentation. The patient reported that his father and brother seemed to have undiagnosed attention and memory problems; anxiety and depression were reported in the extended family. Further, Mr. A's childhood was affected by traumatic abuse that he revealed upon entering treat-

[3] I thank research assistant Sam Hayden for designing this record form.

ment. During his preschool years, he began to be sexually abused by a somewhat older paternal cousin. Mr. A felt dominated by this larger, boisterous boy who engaged him in reciprocal sexual acts.

In school, Mr. A was a capable student who experienced his need to work hard to get good grades as a narcissistic injury. Mostly earning B's, he saw himself as a chronic "underachiever." This tension eased when he attended college in another state (perhaps metaphorically as well as literally) and discovered a strength in English. Despite a lesser talent in math, he pursued a business degree, preparing to join the "home" agency.

Young Adulthood

After graduating from college, Mr. A moved to a more distant city, working in an agency with a mission similar to that pursued by his family. He later returned to his home town as he began to think about "settling down"—marrying and starting a family of his own. He bought a house, developed strong friendships, began to date, and took on a supervisory role within the hometown agency. He drank socially but did not use drugs. His goal upon entering therapy was to feel free enough to marry.

Mr. A's concerns about his ability to establish intimate connections with women were accompanied by fears that his memory problems were undermining his job performance and setting the stage for instances of "humiliation." Sexual concerns intruded on his focus and concentration, and he had become forgetful, not remembering to lock his car and later finding that items within had been stolen. He could not say that he had never before had problems with attention, memory, and organization, but this recent upsurge of scattered thinking felt subjectively different to him.

A troubling instance of work-related forgetting prompted Mr. A's discussion with his therapist about ADHD: Mr. A had been given tickets to an upcoming opera. Unable to use these tickets, he gave the complementary passes to another supervisor 2 days before the event, who in turn gave them to a client. On the day of the performance, Mr. A decided to offer the tickets to a friend, forgetting that he had already given them to his coworker. Unable to find the tickets, he reported them as "not received, or stolen" to the box office. When his coworker's client and family arrived for the performance, the father was placed in handcuffs until the matter was resolved—humiliating for the client, Mr. A, and his coworker. The testing thus focused on identifying the relative contributions and possible intertwining of an unconscious wish and "true" neurocognitive weaknesses.

Clinical Observations and Structure of the Assessment

Mr. A presented for evaluation as a well-groomed man who was dressed in casual business attire. Despite a periodic melancholy, he was an active collaborator in the assessment process, which included an initial 1-hr clinical interview, three testing sessions of 3 hours each, and a feedback session. At the outset, Mr. A sadly mused on the lack of guidance and support available to him as a child. And yet he did not appear frankly depressed. He displayed a full range of affect, exhibited a good level of energy without restlessness or distractibility, spoke at a typical speed and volume, and coherently communicated his thoughts. Aiming for a nuanced exploration of Mr. A's cognitive and emotional landscape, the test battery included measures of intelligence, attention, memory, executive functions, and personality.

Formal Testing: Wisconsin Card Sorting Test-Computer Version 4 (WCST-CV4)

I will describe in detail the response process as it emerged for Mr. A on the Wisconsin and then consider qualitative and quantitative data, integrating the findings within a brief overview of the complete results. Finally, I will share diagnostic impressions and how they informed treatment recommendations.

The Wisconsin, with its status as a neuropsychological measure—but one whose administration involves ambiguous instructions, seating akin to that of psychoanalysis, and an explicit stance of evaluator nonintervention—is perhaps ideally placed at the midpoint of a test battery. It provides a segue between structured tests of the cognitive realm and projective tests[4] that are explicitly designed to assess personality. This draws on the tradition of Rapaport et al. (1968), where tests are organized from most to least structured. Serving as a bridge between neurocognition and personality, the Wisconsin is in a unique position to illuminate the richness offered by carrying a psychoanalytic mindset into the "non-analytic" domain of neuropsychological testing.

Mr. A's response process. Mr. A seated himself comfortably in front of the laptop computer and listened attentively to the standardized set of instructions: "This test is a little unusual because I am not allowed to tell you very much about how to do it . . ." His manner was quietly expectant, and I sensed from him an interested curiosity. At the same time, I had in mind his comments following earlier completion of the intelligence test: "My problem is one of comparison. I feel like less of a person because I always imagine that there are people who will score higher, be worth more."

Mr. A sat forward in his chair, looking carefully at the first stimulus card and clicked the mouse on a key card that matched according to both Form and Number; the computer responded WRONG. He smoothly moved to the next card and made his match based on Color; the computer responded RIGHT. He settled back comfortably in his chair and proceeded to match according to Color for the next nine responses, completing the category (unbeknownst to him). He consistently clicked on the top row of key cards rather than on the row of accumulating response cards and he quickly discerned that he must wait for the banner to recede before it would accept his response, adjusting his pace accordingly. I noted these indications of intact spatial relations and executive functioning as I recorded his responses.

Mr. A made his twelfth response, continuing to click on Color, but at this point the computer responded WRONG. He matched according to color twice more before making a match based on Number; the computer continued to respond WRONG. I sensed his emerging uneasiness. The next response card happened to match on all three dimensions so when Mr. A clicked on the identical key card, he heard RIGHT. He relaxed for a moment—but quickly became tense when subsequent responses again prompted WRONG. In all, while attempting to complete this second cate-

[4] Although, for the sake of historical continuity, I have retained the traditional term *projectives* for tests with ambiguous stimuli that are designed to elicit open-ended responses with no right or wrong answers, there is current interest in relabeling such tests. *Ambiguous-demand performance-based measures* (Bram & Peebles, 2014, p. 36) is one such term that has been proposed.

gory of Form, he responded to 40 stimulus cards without generating 10 consecutive correct responses.

Mr. A's discomfort was becoming palpable. He was saying not a word but sat slumped in his chair; I could feel him alternating between agitation and near paralysis: He would respond quickly to several cards in a row and then sit back, staring almost blankly at the key cards, as if waiting for them to give him the answer. This pattern repeated itself several times. I felt comfortable allowing Mr. A to struggle for a period of time, as I was interested to see whether he might steadfastly persevere and come to appreciate his own resilience, experience an "aha" moment and discover a solution that had been temporarily out of reach, angrily devalue the test itself, helplessly cede control to me, or communicate some other characteristic mode of responding.

At the same time, I was cognizant of how Mr. A's performance deviated from my own internal norms, based on many years of giving this test. In my experience, it is unusual for a patient with average intelligence or above—even those with ADHD or learning disabilities—to have such difficulty with the second category. Typically, it is the shift from Form to Number for the third trial that proves most challenging. And, as I wondered whether neurocognitive deficit or anxiety was the primary impediment at this moment, I thought about how introducing a parameter would offer an opportunity to test my hypothesis that anxiety was primary. Over the years, I have found that for patients ultimately diagnosed with ADHD, simply asking, "How are you making your match?" is often insufficient to organize them. For some, the question is dismissed as an irritating intrusion and they continue to respond impulsively. For others, it allows them to regain a temporary focus that is lost at the next moment of transition. And for yet others, it promotes a tendency to then "insist" on help at moments of cognitive or emotional challenge. In contrast, for those with anxiety, such a parameter often allows them to step back, reflect, restabilize, and problem solve effectively for the remainder of the test.

With Mr. A, not only did I feel that data derived from a standardized but "paralyzed" administration would be of questionable value but also that introducing a parameter would test a hypothesis that was central to the diagnostic question and ultimately inform treatment recommendations. Although maintaining a productive testing alliance was a factor at this midpoint in the assessment, exploring the central referral question was primary. In deciding how to intervene, I considered Mr. A's history, which contained traumatic abuse and subtle sadomasochistic enactments where he alternated between being "victim" and "victimizer." I sought a position that was neither rescuing nor punishing. After he made his 40th response to the Form category without discovering the guiding principle, I said, "Let me ask you to pause for a moment." He did so, glancing back at me with an angst-filled expression. I continued, aiming to convey an interested curiosity, "Tell me, how are you making your match?"

Mr. A turned back to the screen and I could feel him relax. He studied the response and key cards, but with an expression of quiet interest rather than his former blankness. After perhaps 10 seconds, his body became erect. "Oh!" he said, sitting forward slightly. "The shape!" He clicked on the key card corresponding to Form, heard RIGHT, and proceeded to attain 10 consecutive correct responses. Following the next WRONG when the categories shifted again, he experimented with a few responses before comfortably completing Number. As the test cycled through Color, Form, and Number for the second time, he shifted smoothly from one category to the next following a single WRONG cue each time. As the test came to an end, he looked back at me, with a mixture of pride and sadness. "You know," he said, "I don't think I have ADD. I just think I need a little guidance and support to figure things out sometimes."

Quantitative and qualitative findings from the WCST. Quantitative results suggest that Mr. A's executive functions are compromised relative to his intelligence and the normative group. His Full Scale IQ of 108 (70%ile) and his smooth subtest score profile on the Wechsler Adult Intelligence Scale (WAIS-IV; Wechsler, 2008) predict at least average capabilities with respect to abstract reasoning, strategic planning, and self-monitoring on the Wisconsin. Instead, one finds age- and education-corrected scores that cluster 2 standard deviations below his measured intelligence and significantly below the test mean (See Table 1). Only Mr. A's nonperseverative errors are marginally better. Scores in the lower half of the table are grossly intact, reflecting the organizing impact of the parameter and a lesser statistical sensitivity here: Scores on these dimensions indicate only whether performance falls more or less than 1 standard deviation below the mean.

Yet integrating qualitative aspects of Mr. A's performance with these scores yields a more nuanced picture. Consider his microtrauma: the narcissistic injury of never feeling smart enough, and the shame associated with his sense of intellectual inadequacy. Consider also the profound trauma of his ongoing sexual abuse as a child, which likely created a vulnerability to dissociative defenses—seen in Mr. A's shift from initial uneasiness to alternating agitation and paralysis. And yet he maintained his spatial orientation to the top row of key cards and his awareness of the need to allow the banner to recede before responding. In my experience, these qualitative features of the response process are highly vulnerable to disorganization in those with ADHD; therefore, the absence of such lapses offers contraindicating evidence with respect to this diagnosis.

Despite his very low scores, Mr. A seemed to demonstrate strengths in executive functioning that were disrupted by intense

Table 1
WCST Demographically-Corrected Scores for Mr. A

Wisconsin Cart Sorting Test: CV4 (mean standard score = 100, *SD* = 15)			
Trials administered	116		
Total correct	79		
WCST Scores	Standard scores	%ile	T Scores
Total errors	79	8	36
Perseverative responses	72	3	31
Perseverative errors	71	3	31
Nonperseverative errors	89	23	43
Conceptual-level responses	80	9	37
	Raw score		
Categories completed	6 of 6	>16%	
Trials to complete first category	11	>16%	
Failure to maintain set	0	>16%	
Learning to learn	.15	>16%	

anxiety intruding on his reflective function. Indeed, an inverse relationship between reflective function and two variables on the Wisconsin (failure to maintain set and perseverative errors, with the latter representing Mr. A's lowest score) has been empirically demonstrated (Levy et al., 2005). Further, it was a moment of authentic connection between us—the "something more" in a testing context—that allowed Mr. A to regain reflective function. Although he made 37 errors during the course of the test (24 perseverative), all but nine occurred before the parameter was introduced. And he never "lost set" once he had established a category. Indeed, Mr. A completed the remainder of the test capably, restabilizing with a surge of confidence and insight into the type of support he needs to function optimally.

Overview of test results. Complete test results are presented in summary form to retain the paper's focus on the vignette from the Wisconsin, which highlights the value of bringing a psychoanalytic mindset to neuropsychological testing. As noted earlier, Mr. A's intelligence was measured on the cusp of the high average range, and his performance on the Trail Making Test (Reitan, 1986), Stroop Color and Word Test (Golden & Freshwater, 2002), CPT-II (Conners, 2002), and Wide Range Assessment of Memory and Learning (WRAML2; Sheslow & Adams, 2003) met or exceeded expectation relative to his intelligence and age-based test norms. The Wisconsin was the exception to this pattern among evaluator-administered neurocognitive measures.

Given the earlier discussion about the limitations of self-report measures, it is instructive to note that on the five subscales of the Brown ADD Scales (Brown, 1996), a self-report measure assessing various dimensions of the disorder, Mr. A's responses yielded scores more than 3 standard deviations above the mean. His total score of T = 93 (M = 50, SD = 10) resulted in a "highly probable" classification with respect to the likelihood of having ADD. I interpret these findings not as documenting ADHD but as reflecting a "cry for help" profile that captures a patient's subjective experience of profound cognitive disruption, as was the case here.

In the personality domain, on the one self-report measure, Mr. A displayed the same pattern of highly significant elevations across virtually all clinical subscales. His extreme anxiety (T = 85, M = 50, SD = 10) so distorted his profile on the Personality Assessment Inventory (PAI; Morey, 1991) that diagnostic impressions were suspect: Schizophrenia, a primary consideration, was inconsistent with other test data, with my own clinical impressions, and with those of Mr. A's therapist. The PAI, to its credit, noted the "cry for help" profile and urged caution in interpreting results. Indeed, when extremely high scores on self-report measures diverge from other neurocognitive and personality data, they can be conceptualized as capturing a patient's intense subjective distress and potential for destabilization better than they confirm a specific diagnosis.

Given Mr. A's questionable self-report results, the personality assessment relied on projective measures, including chromatic and achromatic drawings (Hammer, 1958), the Early Memory Test (Mayman, 1968), Thematic Apperception Test (TAT; Murray, 1943), and Rorschach. Findings indicated a sensitive and articulate man whose identity development appeared to have been derailed in a manner consistent with early trauma. With evidence of early life experiences that felt too little recognized and modulated, he struggled to develop a separate and integrated sense of self. Mr. A's identity diffusion set the stage for "fuzzy" cognitive states—a sense of knowing and not knowing—and marked dysregulation of affect and impulse. His inability to think and "know" when feeling endangered—too alone or exposed—intruded on attention and memory.

Diagnostic impressions. Diagnostically, test data revealed a narcissistic character structure that left Mr. A vulnerable to fluctuating anxious and depressive states, confused thinking, and a fragile sense of self. As a man, he felt deeply damaged and deficient, especially with regard to his sense of masculinity (with his sole response to Rorschach Card I being "a butterfly with a pinhole through it . . . held down to a piece of paper"). Masculine energy and strength were associated with the twin images of predator and prey. On the one hand, Mr. A portrayed the predator who aggressively scours the landscape to find what he needs for his own survival; on the other hand is the subsequent sense of aloneness and smallness. For example, on Rorschach Card IV, he first asks, "Again, am I to speak openly about everything?" He then describes "a vulture peering down [upper side details] and down here is the landscape [remaining blot with edge details] and obviously if he's a vulture, he's looking for carrion . . . I always picture a vulture looking down because they're always trying to just scour the earth and look for food. And this [center detail] also reminds me of those enormous towers they've been building recently, I can't remember what they're called. The building starts to make me think of Asia and how isolated we get in our own world . . . how separate we all are . . . I was looking at the height of the Eiffel Tower and how it used to be the biggest one and now others are taller than it." Even in the absence of formal scores, one grasps Mr. A's inner world (including "not remembering"), with its polarized representations of self and other.

Feedback session and treatment recommendations. Mr. A returned for a feedback session a month after the testing was completed, with his response process in this final phase of the assessment lending weight to the treatment recommendations. At this moment of heightened anxiety, there emerged a clear instance of the subtly devaluing attitude that could be expected given his earlier idealization. Entering the consulting room from the waiting room where he had been seated for the 10 min prior to the appointed time, Mr. A sat down and reclined slightly, asking me to please get him a drink of water; he explained that his throat might get dry as we discussed his results. Neither wanting to create a narcissistic injury nor masochistically "submit" to his request—and not wanting to take on the interpretive role of the therapist—I offered to show him where the water fountain in the hallway was located, so that he could get a drink before we began our discussion. He appeared a bit startled but accepted this compromise.

Before reviewing the results, I asked Mr. A if there was something that stood out for him from our testing collaboration. He said that he had been struck by how his anxiety had made him unable to focus or think on "that computer test" (the Wisconsin); he began to believe that there was not something "wrong" with his brain but rather that when he feels too unsupported or too dominated, his mind becomes "a confusing mess." He conjectured that this might have something to do with his history of abuse, adding, "And when you didn't just leave me hanging out there but you didn't do it for me, I could think again." I concurred with his analysis, pleased that the parameter had yielded this vivid relational moment that allowed Mr. A to subjectively experience what the test results

confirmed, and which I could then review in experience-near language for him.

Mr. A's unstable inner object relationships created a vulnerability to fluctuations in his ability to adaptively express and contain deeply divided feelings and impulses. And yet he demonstrated the psychological-mindedness and the ability to regain his reflective function in the context of an ongoing therapeutic encounter. Clearly, stimulant medication to treat ADHD was contraindicated given attention and memory problems due to trauma rather than a neurodevelopmental disorder. I felt that treatment focusing on the transference would ultimately regulate Mr. A, allowing him to grapple with the relationships of his polarized inner world—a conceptualization readily grasped by his therapist. With Mr. A, I highlighted the value of exploring moments of interaction with his therapist as a way of coming to know himself, as had occurred during testing. Greater frequency of sessions was expected to allow for both increased intensity and support. Later follow-up with his therapist revealed that Mr. A responded well to this shift in focus and frequency. He had developed a more textured understanding of his inner emotional landscape, had experienced improved attention and memory with fewer enactments, and had achieved successes in his personal life that were at once unexpected and satisfying.

Closing Comments

Bringing a psychoanalytic mindset to neuropsychological testing allows for the depth and complexity of the psychoanalytic tradition to be applied to contemporary assessments. Referral questions focusing on attention and learning problems in children and adults are increasingly common, calling for comprehensive evaluations that can differentiate the intertwined threads. For Mr. A, the assessment provided a window into his functioning, making it possible to confidently rule out ADHD while revealing the personality structure and history of trauma that gave rise to his concerns. Further, integrating a parameter allowed for a "something more" within the testing process—a vivid relational moment that highlighted for both patient and evaluator the intersection between Mr. A's psychodynamics and his problem with remembering and "knowing."

Questions, of course, remain: How does one understand what was enacted in the incident of the "not received, or stolen" tickets? Do these words capture what Mr. A feels was not given, or was taken from him, during his childhood? Is the supervisor to whom he gave the tickets a stand-in for the cousin he wishes to impress and humiliate? Might the handcuffed "father" represent the preoccupied father who unconsciously deserves to be punished? Such questions become grist for the mill of the therapeutic endeavor, better explored and integrated within Mr. A's unfolding psychoanalytic psychotherapy than through the briefer experience of a psychological evaluation.

Yet it is essential to recognize the contribution of a carefully conceived psychodiagnostic test battery for neurocognitive questions. Measures that are selected, administered, and interpreted via the richness of a psychoanalytic mindset provide important diagnostic returns. Through the lens of such an evaluation, key questions are brought into focus, core deficits and conflicts are identified, and thoughtfully conceived recommendations allow the patient and referring therapist to confidently move forward.

References

Allison, J. (1978). Clinical contributions of the Wechsler Adult Intelligence Scale. In B. Wolman (Ed.), *Clinical diagnosis of mental disorders* (pp. 355–392). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4684-2490-4_12

Bornstein, R. F. (2001). The impending death of psychoanalysis. *Psychoanalytic Psychology, 18,* 3–20. http://dx.doi.org/10.1037/0736-9735.18.1.2

Bornstein, R. F. (2005). Reconnecting psychoanalysis to mainstream psychology: Challenges and opportunities. *Psychoanalytic Psychology, 22,* 323–340. http://dx.doi.org/10.1037/0736-9735.22.3.323

Bornstein, R. F. (2010). Psychoanalytic theory as a unifying framework for 21st century personality assessment. *Psychoanalytic Psychology, 27,* 133–152. http://dx.doi.org/10.1037/a0015486

Bram, A. D., & Peebles, M. J. (2014). *Psychological testing that matters: Creating a road map for effective treatment.* Washington, DC: APA Books. http://dx.doi.org/10.1037/14340-000

Bram, A. D., & Yalof, J. (2015). Quantifying complexity: Personality assessment and its relationship with psychoanalysis. *Psychoanalytic Inquiry, 35,* 74–97. http://dx.doi.org/10.1080/07351690.2015.987595

Brown, T. E. (1996). *Brown Attention-Deficit Disorder Scales: For Adolescents and Adults.* San Antonio, TX: PsychCorp/Pearson.

Conners, C. K. (2002). *Conners' Continuous Performance Test II.* North Tonawanda, NY: Multi-Health Systems.

Eissler, K. R. (1953). The effect of the structure of the ego on psychoanalytic technique. *Journal of the American Psychoanalytic Association, 1,* 104–143. http://dx.doi.org/10.1177/000306515300100107

Golden, C. J., & Freshwater, S. M. (2002). *Stroop Color and Word Test: Revised examiner's manual.* Wood Dale, IL: Stoelting.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38,* 404–411. http://dx.doi.org/10.1037/h0059831

Hammer, E. F. (1958). *The clinical application of projective drawings.* Springfield, IL: Charles C. Thomas.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual.* Lutz, FL: Psychological Assessment Resources.

Leak, S., & Hayden, S. (2015). *Shame and guilt in dyslexia and attention-deficit disorder: An integration of neurocognitive and psychodynamic perspectives.* Manuscript under review.

Levy, K. N., Meehan, K. B., Reynoso, J. S., Lenzenwegger, M. F., Clarkin, J. F., & Kernberg, O. F. (2005). Abstracts of the 2005 poster session of the American Psychoanalytic Association Winter Meeting: The relation of reflective function to neurocognitive functioning in patients with borderline personality disorder. *Journal of the American Psychoanalytic Association, 53,* 1305–1308.

Lezak, M. D. (1983). *Neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Mayman, M. (1968). Early memories and character structure. *Journal of Projective Techniques & Personality Assessment, 32,* 303–316. http://dx.doi.org/10.1080/0091651X.1968.10120488

Meersand, P. (2011). Psychological testing and the analytically trained child psychologist. *Psychoanalytic Psychology, 28,* 117–131. http://dx.doi.org/10.1037/a0021043

Morey, L. C. (1991). *Personality Assessment Inventory.* Odessa, FL: PAR.

Murray, H. A. (1943). *Thematic Apperception Test.* Cambridge, MA: Harvard University Press.

Rapaport, D., Gill, M., & Schafer, R. (1968). *Diagnostic psychological testing* (Rev. ed.). Madison, CT: International Universities Press.

Reitan, R. M. (1986). *Trail Making Test: Manual for administration and scoring.* Tucson, AZ: Reitan Neuropsychological Laboratory.

Rothstein, A., Benjamin, L., Crosby, M., & Eisenstadt, K. (1988). *Learning disorders: An integration of neuropsychological and psychoanalytic considerations.* Madison, CT: International Universities Press.

Rothstein, A., & Glenn, J. (1999). *Learning disabilities and psychic conflict: A psychoanalytic casebook.* Madison, CT: International Universities Press.

Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing.* New York, NY: Grune & Stratton.

Sheslow, D., & Adams, W. (2003). *WRAML2: Wide Range Assessment of Memory and Learning* (2nd ed.). Lutz, FL: PAR.

Stern, D. N., Sander, L. W., Nahum, J. P., Harrison, A. M., Lyons-Ruth, K., Morgan, A. C., . . . Tronick, E. Z. (1998). Non-interpretive mechanisms in psychoanalytic therapy. The 'something more' than interpretation. *The International Journal of Psychoanalysis, 79,* 903–921.

Strauss, E., Sherman, M. S., & Spreen, O. S. (2006). *A compendium of neuropsychological tests* (3rd ed.). Oxford, UK: Oxford University Press.

Sugarman, A., & Kanner, K. (2000). The contribution of psychoanalytic theory to psychological testing. *Psychoanalytic Psychology, 17,* 3–23. http://dx.doi.org/10.1037/0736-9735.17.1.3

Tuber, S. (2012). *Understanding personality through projective testing.* New York, NY: Jason Aronson.

Wechsler, D. (2008). *WAIS-IV administration and scoring manual.* San Antonio, TX: NCS Pearson.

Yalof, J. (2006). Case illustration of a boy with nonverbal learning disorder and Asperger's features: Neuropsychological and personality assessment. *Journal of Personality Assessment, 87,* 15–34. http://dx.doi.org/10.1207/s15327752jpa8701_02

Psychology, Public Policy, and Law
2018, Vol. 24, No. 1, 11–23

1076-8971/18/$12.00 http://dx.doi.org/10.1037/law0000140

A Hierarchy of Expert Performance Applied to Forensic Psychological Assessments

Itiel E. Dror
University College London

Daniel C. Murrie
University of Virginia

Experts in forensic psychology must make skilled observations and conclusions, minimally compromised by bias, in order to try and provide reliable and accurate conclusions to the courts. But the field has little data revealing how well forensic psychologists actually perform these tasks, in part because there has been no clear framework for systematic research of their expertise. Therefore, we consider forensic psychological assessments in light of Dror's (2016) Hierarchy of Expert Performance (HEP). HEP addresses reliability and biasability, both within and between experts, at the levels of observations and conclusions. Applying this framework to forensic psychological assessments reveals a few domains in which there are some meaningful data, particularly addressing reliability between experts in certain types of forensic assessments. But applying HEP reveals more domains in which we lack data addressing fundamental aspects of expert performance, such as reliability at the level of observations, and reliability and biasability *within* experts. Understanding these strengths and gaps in forensic assessment research should guide testimony of forensic psychologists, policies around forensic assessment, and further research in forensic assessment.

Keywords: bias, reliability, adversarial allegiance, contextual effects, forensic assessment

Expert performance can be characterized and measured in a variety of ways. When it comes to decision making, a basic measurement is accuracy. In other words, are expert decisions correct, and do they reflect the ground truth? But there are many domains—including many areas in forensic science and forensic psychology—wherein the "real" answer is never known. For example, who actually committed the crime may be unknown, or the defendant's actual mental state at the time of the crime may be unknown. Because ground truth is unknown in real cases, we cannot always directly measure the accuracy of forensic experts. Given this problem, controlled research is needed to help ascertain accuracy less directly, by determining the components involved in forensic decision making. Then we can experimentally isolate and measure them, with a view that such understandings will reveal strengths and weaknesses, and then inform strategies and policies to improve the work of forensic experts.

This article was published Online First September 21, 2017.

Itiel E. Dror, University College London; Daniel C. Murrie, Institute of Law, Psychiatry, and Public Policy, University of Virginia.

We thank John Monahan for his helpful comments on a draft of this article. Some content of this article was presented at the annual conference of the American Psychology-Law Society (Seattle, Washington), the Risk and Recovery Conference (Hamilton, Ontario), and the British Psychological Society annual forensic conference (Bristol, United Kingdom), all in 2017. Each author made equal contributions to this article, so authorship order was determined by coin flip.

Correspondence concerning this article should be addressed to Daniel C. Murrie, Institute of Law, Psychiatry, & Public Policy, University of Virginia Box 800660, Charlottesville, VA 22908-0660. E-mail: murrie@virginia.edu

A Hierarchy of Expert Performance

Two basic properties of decision making are biasability and reliability (Dror, 2016; Dror & Rosenthal, 2008). *Biasability*, within the forensic science and legal community, refers to the potential effects of irrelevant contextual information and other biases that may impact the decision. For example, would a criminal defendant's race influence forensic evaluators' decisions (as it seems to among other decision-makers in the justice system; Mitchell, Haw, Pfeifer, & Meissner, 2005; Smalarz, Madon, Tang, Guyll, & Buck, 2016)? *Reliability* refers to the consistency, reproducibility, or repeatability of decisions, regardless of bias. For example, would different forensic evaluators reach the same conclusions about a defendant's legal sanity when they review and examine the same collateral records and recorded interview of the defendant? Reliability and biasability are distinct concepts, but both contribute to variability in decision making.

Without considering and teasing apart the different elements underpinning expert performance, such as reliability and biasability, it is hard to properly quantify expert performance, particularly because there are no parameters to research. Organizing different elements in expert performance enables us to understand the different aspects of expert performance and how they relate to one another. Reliability and biasability are often lumped together, and therefore variability in decisions may not be correctly attributed to the biasing effects or to the reliability per se. Furthermore, teasing apart the different components of decision making allows us to identify gaps in the literature and to prescribe further research, as well as to direct and focus policies on specific problems, where needed. For example, policies need not address basic reliability (e.g., by requiring multiple assessments) if biasability is the major contributor to

the variance; conversely, policies need not address biasability (e.g., by requiring a neutral, court-appointed expert) if simple unreliability is the underlying problem.

A second distinction, in addition to reliability and biasability, is to quantify experts' performance relative to other experts (*between experts*, or interexpert performance) versus experts' performance relative to themselves (*within experts*, or intraexpert performance). For example, a between-expert study might examine whether several different fingerprint examiners identify the same features in the same fingerprint mark, whereas a within-expert performance study might examine whether the same fingerprint examiner will identify the same features in the same fingerprint when it is presented multiple times at different occasions (e.g., Dror, et al., 2011).

Within-expert performance is perhaps the most basic and essential level of expertise. When experts vary in their performance among one another (between-experts), this variability can be attributed to individual differences (e.g., different philosophies and ideologies, different training and experience, different subjective decision thresholds, different eyesight, different risk tolerance, and a variety of factors that make experts different from one another). However, if an individual expert cannot be consistent with himself—that is, if he or she cannot draw the same observations and conclusions from the same data—this unreliability cannot be attributed to individual differences. Thus within, intraexpert, performance measurements are a more basic metric, and foundational to expert performance.

A third distinction in studying expert decision making is the distinction between observations and conclusions. Conclusions depend on assessment and interpretation of observations. Therefore, to understand decisions, one must be able to distinguish performance in interpreting observations (i.e., drawing conclusions) versus performance in actually *making* the original observations (Dror, 2016). Lumping these together obscures the initial observational performance and may be misleading because observations underpin the resulting conclusions.

In the medical domain the distinction between observation and conclusion is made clear and explicit in the SBAR (Situation, Background, Assessment, and Recommendation) protocol (e.g., Thomas, Bertram, & Johnson, 2009; Wacogne & Diwakar, 2010). The Situation and Background focuses on observations (such as, patient heart rate is 140, patient has a history of heart attacks, etc.), whereas the Assessment and Recommendation focuses on the conclusions based on the observations (such as, patient is having a heart attack, patient should be provided with oxygen, etc.). Similarly, the "SOAP" method (Subjective, Objective, Assessment, Plan) guides clinicians to work from observations to conclusions (Weed, 1970). "Subjective" and "Objective" refer to observations about the patient's presentation, whereas "Assessment" and "Plan" are conclusions based on the initial observations.

Research from the forensic sciences illustrates the crucial distinction between observations and conclusions. For example, imagine two fingerprint experts reach different decisions: one expert concludes with high confidence that the two prints 'match' (i.e., come from the same source) whereas the other expert concludes with high confidence that they do not match. It may be that both examiners observed the same data in the fingerprints (the minutia; see Figure 1), but nevertheless reached opposing conclusions because they use different similarity thresholds for concluding a 'match' (e.g., one examiner requires a minimum of 12 matching points, whereas the other requires at least 13). Alternatively, it may be that the two examiners used identical thresholds to reach a conclusion of a 'match' (e.g., 12), but reached different decisions because they observed different data in the fingerprints (e.g., one examiner observed the 12 minutia needed to call a 'match,' whereas the other only observed 9) (Dror et al., 2011). Understanding this distinction is crucial for designing interventions to increase reliability; for example, do the experts need better training in observing fingerprint minutia, better training in drawing conclusions, or both?

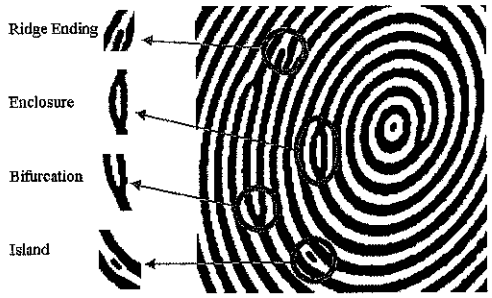

Figure 1. Different characteristics (minutia) that may be present in a fingermark.

This between-expert (interexpert) example illustrates why it is important to tease apart observations from conclusions, but the distinction can equally apply to within-expert (intraexpert) performance. If the same expert makes a different decision when the same fingerprints are presented multiple times at different occasions, this can happen because the expert used different decision rules at the different times, or it can be because the expert observed different data at the different times. Here, when the variance is within-expert, the interventions will be different than those for between-expert variance (hence the importance of distinguishing between- and within-expert variance); for example, has change in eyesight underpinned observing different minutia (and if so, shall there be a policy for yearly eyesight testing of fingerprint examiners)? Or, do changing work environment and pressures need addressing because they underpin variance in drawing conclusions? Of course, it can be that both observation and conclusion differences contribute to the differences in decisions, and the distinction is not always clear. Nevertheless, it is important to tease them apart as much as possible.

Using these three dimensions of: a) reliability and biasability; b) within experts and between experts; and c) observations and conclusions, Dror (2016) suggested an 8-level Hierarchy of Expert Performance (HEP). As illustrated in Figure 2, at the bottom of the HEP is the most basic measurement of expert performance: reliability within expert observation; that is, how consistent an expert is with herself in what she observes. The question and quantification at Level 1 is the extent to which an expert will observe the same things when presented with the same data. For example, fingerprint experts, presented with the same print at Time 1 and Time 2, will often observe different features (see Dror et al., 2011, for details). Level 2 of HEP remains in the observation stage and still focuses on reliability, but at this level the measurement is differences in performance between experts, rather than within

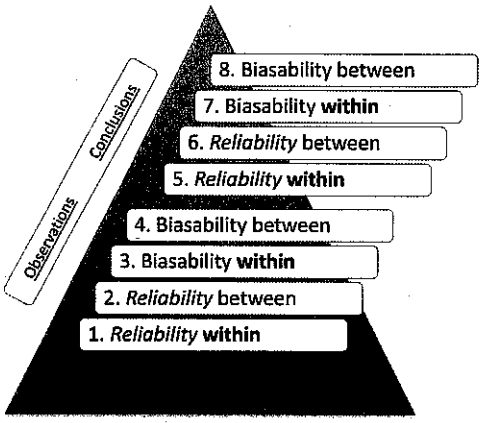

Figure 2. Dror's (2016) HEP: Hierarchy of Expert Performance. See the online article for the color version of this figure.

experts. Levels 3 and 4 examine biasability in observations—Level 3 within experts and Level 4 between experts. These levels address the impact of irrelevant contextual information on observations of data. For example, will irrelevant context (such as the nature of the crime or the suspect) influence what experts observe in the fingerprint evidence (e.g., Earwaker, Morgan, Harris, & Hall, 2015)?

While Levels 1–4 focus on observations, Levels 5–8 of HEP focus on the conclusions. Following the structure of Levels 1–4, Levels 5–8 address quantification of reliability within experts (Level 5), reliability between experts (Level 6), biasability within experts (Level 7), and biasability between experts (Level 8).

Expert Performance in Forensic Psychology and Psychiatry: Identifying Research Needs and Policy Implications

Like other forensic science practitioners, forensic psychologists and psychiatrists are trained experts, often assigned to review complex case materials, interview criminal defendants or civil litigants, and provide expert opinions to assist the court. Also like other forensic science experts, their services usually culminate in formal written reports or expert testimony in which they present conclusions to assist a judge or jury in reaching a verdict. To the extent that their conclusions are reliable and resistant to bias, the justice system can have greater faith in them. But to the extent that these conclusions are unreliable or biased, they are at best unhelpful, and at worst misleading, to the justice system's goals of administering justice with accuracy and equity.

Despite their similarities to other forensic scientists, forensic psychologists and psychiatrists have escaped much of the public and government scrutiny that other forensic science domains have received. The National Research Council (NRC, 2009) and the President's Council of Advisors on Science and Technology (PCAST, 2016) reviewed the state of forensic science, covering a wide range of disciplines including analyses of DNA, fingerprints, hair, tool marks, bite marks, and firearms. Both government councils concluded that the reliability of many forensic techniques is unknown and that forensic scientists are prone to a variety of contextual biases.

Wide-scale calls for reform prompted vigorous national efforts, such the National Institute of Standards and Technology's (NIST) efforts to develop best practices and standards in the forensic sciences, and the formation of a National Commission on Forensic Science to develop policies. Both entities include specialized Human Factors groups to address human decision making and bias, and both have produced a wide range of policies to increase reliability and reduce bias in the forensic sciences (e.g., NCFS, 2015). NIST has even provided substantial funding to improve the scientific foundations of forensic science (National Institute of Standards and Technology, 2016). *None* of the recent wide-scale reviews or reform efforts have addressed forensic psychology or psychiatry. But regardless of whether authorities expand their scrutiny (and funding) of forensic sciences to include forensic psychology, their efforts evoke similar questions about forensic psychology and psychiatry (Guarnera, Murrie, & Boccaccini, 2017; Heilbrun & Brooks, 2010; Murrie, Boccaccini, Guarnera, & Rufino, 2013). Likewise, all relevant ethical and professional standards (e.g., AERA, APA, & NCME, 2014; APA, 2002, 2013) suggest the field has a duty to examine and optimize the reliability and objectivity of our methods, regardless of external scrutiny.

Just as Dror's (2016) hierarchy of expert performance (HEP) provided both theoretical and practical benefits to understanding expert performance in the forensic sciences, applying this HEP to forensic mental health evaluations allows us to better understand the expertise and performance of forensic psychologists and psychiatrists. It should help identify areas of strength and weakness, as well as areas in which we simply lack adequate data about evaluators' expert decision making. Following Dror's (2016) HEP in forensic science, we will work from the top of the hierarchy (Level 8) to the bottom (Level 1), beginning with the considerations that appear most visible (e.g., biasability between expert conclusions) and working down to the most basic questions of reliability in observations within experts, which have been least observed or researched. A primary goal in this paper is to identify where empirical data is lacking—that is, where we know little about the performance of forensic psychologists providing expert services to the justice system—and prescribe research that will address these gaps and ultimately improve expert decision making.

In performing this literature review, we used several strategies. We performed standard searches of online research databases using many variations of terms related to bias, reliability, and forensic psychology. We reviewed authoritative texts (e.g., Melton, Petrila, Poythress, & Slobogin, 2007; Packer, 2009; Zapf & Roesch, 2009) and meta-analyses (e.g., Guarnera & Murrie, 2017) addressing forensic evaluation. Upon locating appropriate studies, we reviewed reference lists and ran "cited by" searches to seek additional relevant studies. Finally, we distributed drafts of this manuscript to authorities in forensic psychology—particularly authorities in the subdisciplines with which we were least familiar—asking that they notify us of any potentially relevant research we may have missed.

Dror's (2016) Hierarchy of Expert Performance Applied to Forensic Psychology

8. Biasability Between Experts' Conclusions

Are different experts, considering identical data, biased by irrelevant contextual information? In forensic science, irrelevant contextual information may include whether a suspect confessed to the crime, whether other lines of evidence suggest he is the culprit, and so forth. Researchers have identified the biasing effect of contextually irrelevant information in several ways. For example, Dror and Hampikian (2011) examined whether irrelevant contextual information that implicated a suspect in a sexual assault biased the conclusions of DNA experts. The examiners who were exposed to the biasing information concluded that he could not be excluded from being a contributor to the DNA mixture, whereas most examiners (16 out of 17) who were not exposed to the biasing information did not reach the same conclusion.

In forensic psychology and psychiatry, experts likely encounter a variety of irrelevant contextual information (e.g., case details that are irrelevant and go beyond the specific referral question for forensic evaluation and beyond the expertise of the evaluator). The field has certainly not considered or identified task-irrelevant information in the ways that other forensic sciences have (see, e.g., NCFS, 2015).[1] However, one piece of contextual information that is almost always irrelevant is the "side," or party, requesting an expert opinion.[2] Forensic mental health professionals are to strive for accuracy and neutrality, impartial to the side that retained them (APA, 2013).[3] But judges, legal scholars, and even laypersons have long lamented that forensic experts appear biased by the side that retained them (e.g., Foster, 1897; Hand, 1901; Wigmore, 1923). Likewise, a series of field studies strongly suggest that at least some forensic mental health experts may be vulnerable to *adversarial allegiance*, a bias toward reaching conclusions that favor the party retaining their services (Murrie & Boccaccini, 2015; Murrie, Boccaccini, Johnson, & Janke, 2008; Murrie et al., 2009).

Compelling evidence of biasability between expert conclusions comes from an experimental study in which researchers recruited over 100 practicing, doctoral-level forensic psychologists and psychiatrists and led them to believe they were performing a formal, large-scale forensic consultation (Murrie et al., 2013). These forensic experts were—unbeknownst to them—randomly assigned to either believe that they were paid by the public defender service or the special prosecution unit. These participants met with an attorney who posed as leading either the public defender service or the specialized prosecution unit, and requested that they score particular risk instruments based on extensive offender records (a type of consultation not uncommon in forensic practice). Each participant scored the same four case files, and each file was authentic (i.e., from an actual case), including extensive records (e.g., police, court, correctional, mental health) typical of those evaluators use to score risk instruments in forensic evaluations. Thus, participating forensic experts were able to score the same two commonly used risk instruments that served as the metrics for bias in earlier field studies (Murrie et al., 2008, 2009): that is, the PCL-R (Hare, 2003) and the Static-99R (Helmus et al., 2012).

Overall, the risk measure scores assigned by the experts showed a clear pattern of differences (i.e., experts who thought they were working for the prosecution assigned significantly higher scores, and those who thought they were working for the defense assigned lower scores), revealing biasability between experts as a function of the side they believed retained them. Allegiance effects were stronger for the PCL-R—a measure that requires more subjective clinical judgment—than for the Static-99R, a more structured measure that permits less judgment.

These experimental results provide strong evidence that even scores on ostensibly objective forensic instruments can be compromised by bias (Murrie et al., 2013). To be clear, there was considerable variability in scores even among experts assigned to the same side, a form of poor reliability between experts (distinct from biasability), and not all experts demonstrated allegiance effects. But finding such evidence of allegiance in a context where all other possible explanations have been experimentally controlled suggests that adversarial allegiance is a significant biasing influence among some experts. Put simply, the "side" retaining an expert is one piece of biasing, but irrelevant, contextual information that can cause experts to reach different conclusions.

Beyond this one type of irrelevant biasing information, the field has little data on other types of task-irrelevant contextual information that may bias experts. But many questions about contextual bias are important for practice and policy. For example, an evaluator's role and workplace are theoretically irrelevant to a defendant's true adjudicative competence. But might a community-based evaluator reach a different decision about the competence of a volatile and disruptive defendant than the hospital-based evaluator who would also be tasked with providing treatment or competence restoration services if the defendant is found incompetent and hospitalized? One prominent judge lecturing to psychologists mused, "I wonder how many clinical evaluations that inform decisions about [civil] commitment are influenced by such extraneous considerations as the amount of bed space in the receiving institution" (Bazelon, 1982).

Other potentially biasing information could be fairly easily researched between evaluators (or even within evaluators; see

[1] We acknowledge that determining what is task-irrelevant information is not always clear in forensic science and can be even less clear in forensic psychology. For example, crime details are highly relevant when evaluating legal sanity (a defendant's mental state at the time of the crime), but usually much less relevant when evaluating a defendant's trial competence. A full discussion of this important issue is beyond the scope of this paper, though we note that some contextual information should almost always be irrelevant (e.g., the side retaining the evaluation, the fee for the evaluation, the census pressures in a state hospital, etc.), whereas some (e.g., defendant's sexual behaviors and interests, defendant's family relationships, etc.) may be relevant only to some referral questions and not others.

[2] To be clear, there are a few contexts in which a forensic evaluator's role may properly differ depending on the side retaining that evaluator. For example, Virginia statute directs the mitigation evaluator in a capital sentencing evaluation to actively *assist the defense* in presenting mitigation evidence (Virginia Code § 19.2-264.3:1). But in most situations, the evaluation task, scope, or conclusions should not differ based on the party requesting services.

[3] In the U.S., the Federal Rule of Evidence 702 requirement that "the expert has reliably applied the principles and methods to the facts of the case" suggests an expectation of objectivity. More than 40 states have adopted this federal rule in their state evidence codes. In the United Kingdom, The Criminal Procedures Rule 33.2(1)(a) clearly and explicitly states that the expert's duty to the court must be "objective and unbiased;" bias toward the retaining party is a clear violation.

below) using popular research strategies such as vignettes or case descriptions. For example, would a psychologist reach the same conclusions about a defendant's trial competence if that defendant was described as amiable and nonviolent versus psychopathic and pedophilic (paraphilias and personality disorders are usually irrelevant to trial competence)? Could the race, ethnicity, religion, or sexual orientation of the defendant impact evaluator conclusions if all other case details were identical (see Smalarz et al., 2016 for a similar study within forensic science evidence)? Studies of between-expert biasability may be among the easiest to explore with common social science research methods.

7. Biasability Within Experts' Conclusions

Would the *same* expert reach the same (or different) conclusions when an identical case is presented within a different, irrelevant, biasing context? Whereas Level 8 in the HEP (see Figure 2) deals with differences *between* experts, Level 7 deals with intraexpert, *within*-expert biasability in conclusions. Research in the forensic sciences provides examples of biasability within the same expert's conclusions. For example, the same fingerprint experts did not always reach the same conclusions when the same fingerprints were presented to them on different occasions within different irrelevant contexts (Dror & Rosenthal, 2008). The level of variability in their conclusions depended on a number of factors: the strength of the biasing irrelevant contextual information (which can be relatively weak, such as "the detective believes the person is guilty," or can be relatively strong, such as "the suspect was in custody in jail when the crime was committed"), and the difficulty of the decision (when decisions are more complex and difficult, there is more leeway for bias to influence the conclusions).

In forensic psychology, we know of no comparable research. Within-expert studies are generally harder to conduct than between-expert studies and probably even more difficult to conduct in forensic psychology than in forensic science. Whereas a forensic science expert may not so easily recognize that she is examining the same fingerprints or gunshot residue sample she examined one year earlier, a forensic psychologist would more likely recognize she is examining the same defendant for the same referral question. Nevertheless, these types of studies are critical if we are to explore, understand, and minimize potential biases. Researchers could use methods that do not require repeated exposure to an actual defendant (e.g., presenting vignettes, case summaries, case referrals, or results of psychological tests). For example, would a forensic psychiatrist respond the same way to an attorney's case description if that attorney mentioned a minimal pay rate typical for evaluations of indigent defendants versus a lucrative hourly rate with the promise of more lucrative referrals? Does a psychologist who works in a psychiatric hospital make the same decision about a defendant's restoration to competence when the hospital census is high and hospital beds are scarce as compared to times when there are fewer space constraints? Though such studies may be more challenging to conduct in forensic psychology than in forensic science, the underlying questions are ripe for study and important to explore.

6. Reliability Between Experts' Conclusions

Even without biasing contextual information, will experts examining the same information reach the same conclusions? In forensic science, studies of the basic reliability among experts show that even DNA and fingerprint experts will reach a spectrum of different (and conflicting) conclusions when they examine the same evidence, even absent any irrelevant biasing information about the suspect's likely guilt (e.g., Coble, 2015; Dror & Hampikian, 2011; Dror & Rosenthal, 2008).

Poor reliability between experts threatens the goals of accurate, equitable justice. Stated bluntly, poor between-expert reliability means that whether a suspect goes to prison or not may depend on the 'luck of the draw' as to which expert examines the evidence. Furthermore, since forensic evidence is rarely contested in court, when a fingerprint expert testifying that the fingerprint from the crime scene matches the suspect, it is highly incriminating and often results in a conviction. However, if a different fingerprint expert would have examined the same fingerprints—purely by chance lab assignment procedures—that expert may have not concluded that the prints matched. Hence the dire consequence of a dangerous mix: forensic science evidence is very powerful in court, but different forensic science experts may reach different conclusions. Indeed, many wrongful convictions have relied on confident, uncontested expert conclusions about forensic evidence (Garrett & Neufeld, 2009).

In forensic psychology and psychiatry, several studies provide data about the reliability among expert conclusions (the between, interexpert performance). These data may take the form of reliability estimates for clinicians administering the same instrument (e.g., Boccaccini et al., 2012; Otto et al., 1998; Rogers, Jackson, Sewell, Tillbrook, & Martin, 2003) or clinicians performing the same forensic assessment (e.g., Gowensmith, Sessarego, et al., 2017). Also relevant are studies of whether clinicians assign the same diagnosis to the same individual or case description. For example, in medical research on psychiatric diagnosis, agreement has been poor ($\kappa < .50$) among clinicians who use unstandardized procedures to assign diagnoses (Aboraya, Rankin, France, El-Missiry, & John, 2006; Spitzer & Fleiss, 1974), and the minimal research on diagnosis in forensic evaluations finds similarly poor agreement (Gowensmith, Murrie, Boccaccini, & McNichols, 2017). Recently, scholars have described an important distinction between reliability under the optimal conditions in formal research studies versus "field reliability" among real-world practicing clinicians performing their routine duties within the conditions typical of their work (Edens & Boccaccini, 2017; Wood, Nezworski, & Stejskal, 1996).

A recent review of the field reliability of the most common forensic evaluations—adjudicative competence and legal sanity—identified 59 studies that purported to address the reliability of competence or sanity opinions, but only 8 (for sanity) and 9 (for competence) actually addressed the reliability among practicing forensic evaluators performing real evaluations (Guarnera & Murrie, 2017). These reported pairwise percent-agreement rates ranged from 57% to 100%, and kappa values ranged from .28 (poor) to 1.0 (perfect).

The studies that best shed light on the routine field-reliability among forensic evaluators are from Hawaii, which historically required (Hawaii Revised Statutes, 2014) three independent evaluations for all felony defendants referred for competence or sanity evaluations. Because the evaluators are relatively independent (court-appointed, not retained by the prosecution or defense) they are less vulnerable to the biasing effect of adversarial allegiance.

Thus, Hawaii provides a "natural experiment" for studying field reliability, without the obvious confounds that bedevil other field studies. Regarding adjudicative competence, a review of 216 Hawaii felony defendants referred for evaluation revealed that the three independent evaluators reached different conclusions in 29% of the cases (Gowensmith, Murrie, & Boccaccini, 2012). Regarding legal sanity, a review of 165 defendants revealed that three independent experts reached different conclusions in 45% of the cases (Gowensmith et al., 2013). Finally, regarding evaluations of whether or not a patient who had been hospitalized as not guilty by reason of insanity (NGRI) was ready for conditional release—a legal question less well-defined in statute than competence or sanity—three independent experts reached different conclusions in 47% of the cases (Gowensmith et al., 2017). Hence, field reliability research suggests that expert reliability tends to be modest, even with common evaluations and with court-appointed forensic experts.

Although reliability among experts is usually measured by examining whether experts reach the same conclusion in the same case, it is also possible to explore reliability by examining other more detailed and sensitive measures within a case, or to examine *patterns* of findings across cases from the same referral stream. One example of more detailed or sensitive measures within a case might be confidence levels (e.g., Douglas & Ogloff, 2003). Even if experts reach the same conclusion, they may have different levels of confidence in their judgments, a distinction that is theoretically and practically important. Theoretically, confidence level serves as a more sensitive measure to understand the decision processes (much like researchers use response time as a more sensitive measure to understand differences even when participants give the same response). Practically, confidence is important because a judge or jury may weigh expert opinions based on the testimony and how strongly experts present their conclusions, which depend on the experts' own confidence in their conclusions. Two experts arriving at the same conclusion may nevertheless convey it quite differently because they differ in the confidence they have in the conclusion. In short, reliability can be examined with more sensitive measures beyond overall conclusion.

Another focus in examining reliability (or biasability) might be examining *how* evaluators communicate findings. For example, evaluators presenting the results of certain violence risk assessments may choose between: sharing numerical estimates in frequency or probability formats (Slovic, Monahan, & MacGregor, 2000), describing potential outcomes in "vivid" or "pallid" terms (Monahan et al., 2002), or describing risk factors in "packed" or "unpacked" format (Scurich, Monahan, & John, 2012). Research reveals that each of these decisions about risk communication substantially influences how decision-makers interpret the risk message, even when that message is substantively identical. Thus risk communication strategies may be another focus of reliability (or biasability) research.

Reliability patterns *across* cases is another possible measurement. If evaluators are generally reliable with one another, evaluators working in the same context with the same referral stream should generally display similar patterns of findings across cases. For example, they might find similar percentages of defendants incompetent or insane, and they might, overall, assign similar mean scores on the same instrument, at least *if* all examinees they evaluate are selected randomly from the same population and there is a sufficient number of cases.

In a study that best illustrates this kind of reliability research, Boccaccini, Turner, and Murrie (2008) examined 20 forensic mental health experts who had contracted with the state of Texas to perform screening evaluations of offenders—including administering and scoring Hare's (2003) Psychopathy Checklist-Revised—as part of specialized "sexually violent predator" laws. All experts examined offenders from the same correctional system, referred through the same office, which made efforts to ensure that offenders were assigned to experts in an essentially random manner. In other words, all experts worked for the same entity, examined offenders from the same "referral stream," and should not have seen systematically different samples of offenders. Nevertheless, results revealed that in the 321 cases they examined, the experts differed drastically in the average PCL-R score they assigned across the cases they saw. Indeed, 34% of the variance in PCL-R total scores was attributable to differences between experts, rather than differences in the offenders they evaluated. For example, some evaluators assigned average scores around or above 30 (indicating highly psychopathic personality) whereas some assigned average scores as low as 8 or 18 (scores on the PCL-R can range from 0 to 40, and research reveals average scores of 22–23 in correctional settings). Similar research suggests that experts differ in the proportion of criminal defendants they conclude are incompetent to stand trial (Murrie, Boccaccini, Zapf, Warren, & Henderson, 2008) or not guilty by reason of insanity (Murrie & Warren, 2005), though in most field studies, it is difficult to exclude all confounds (beyond the expert) that might contribute to this variability.

Overall, most research addressing expert performance in forensic psychology tends to fall within Level 6 in the HEP, *reliability between expert conclusions* (see Figure 2). But even in this Level 6 where there are the most studies (e.g., competence and sanity evaluations, scoring forensic assessment instruments), data are sparse and lack some details necessary to answer important practical questions (Guarnera & Murrie, 2017). For example, the forensic psychology literature has virtually no Level 6 data addressing reliability of forensic psychological evaluations in civil litigation or death penalty cases, where the stakes are highest. There is also lack of reliability data for commonplace civil evaluations like those addressing civil commitment or parenting capacity—situations in which courts may restrict rights (or tolerate risk of great harm) based on the opinion of an evaluator. Better reliability data may reveal training needs and may help inform policies such as those that allow or assign multiple evaluations (as in Hawaii) or allow for second-opinion evaluations (as in other jurisdictions).

With further insights on the nature and factors that affect reliability, specific policies and measures can be developed to increase reliability in forensic psychology. In the forensic sciences, more extensive reliability research has produced policies and tools that help quantify and calibrate expert performance, for example, Fingerprint Analyses Consistency Tester (FACT; Dror et al., 2011). We acknowledge that the challenges to develop such policies and tools in forensic psychology may be greater than those in forensic science; for example, some behavioral evidence may be harder to quantify than physical evidence. However, the aspects of forensic psychology that make it more difficult to develop such tools are

the same aspects that are likely to contribute to variability in expert performance, and hence the need for interventions to reduce them.

5. Reliability Within Experts' Conclusions

Level 5 in the HEP examines the reliability of conclusions *within,* rather than *between* experts. This within-expert level is a more basic measure of reliability, in that we would expect an expert to reach the same conclusion if considering the same data repeatedly (even if not reaching the same conclusions as other experts). Forensic science researchers examining this type of reliability found, for example, that even the same fingerprint expert examining the same pair of prints will not reach the same conclusions 10% of the time (Ulery, Hicklin, Buscaglia, & Roberts, 2012).

In forensic psychology and psychiatry, we know of *no* analogous research or any data that examines this aspect of expert performance. As we saw in Level 7 of the HEP, *within*-expert studies are much more challenging and difficult to conduct in general, and even more so in the mental health fields. As Kraemer, Kupfer, Clarke, Narrow, and Regier (2012) observed when reviewing diagnostic reliability in psychiatry, "Intra-rater reliability requires that the same rater be asked to "blindly" review the same patient material two or more times . . . Intrarater reliability is almost never assessed for psychiatric diagnosis because it is difficult to ensure blinding of two diagnoses by the same clinician viewing, for example, the same diagnostic interview" (p. 14). Forensic mental health experts rarely examine exactly the same case data (the way a forensic scientist might reexamine the same evidence), and even if they examine the same defendant, the defendant may have changed. For example, an expert might conclude that a defendant is not competent to stand trial but then evaluate her again after she receives competence restoration treatment and conclude that she is competent. Likewise, an expert might evaluate an individual and conclude that he presents little risk of violence, but evaluate him again later, after changes in dynamic risk factors such as psychosis and substance abuse, and conclude that he is at higher risk for violence (see generally, Douglas & Skeem, 2005). In neither scenario would we consider the experts' differing opinions to be unreliable; the changing conclusions may be appropriate because the case has changed. However, the critical question is: all things being equal, would the same examiner reach the same conclusions from the same data? That is a most basic form of reliability underlying forensic conclusions, and there has been no research examining this fundamental aspect of forensic psychology expertise.

Within-expert reliability is so fundamental for expertise that it should become a priority for future research. The challenge for researchers studying the reliability within expert conclusions would be to present precisely the same case data to an expert at different points in time (without the expert recalling having seen it before). Although it may be nearly impossible to do such studies with live defendants, it may be quite feasible with case files or test results, much like sharing fingerprints or other forensic evidence with forensic science experts.

1–4. The Observational Levels

Whereas levels 5 through 8 of the HEP (above) address expert conclusions, levels 1 through 4 address the underlying, more fundamental observations on which conclusions are based (see Figure 2). Failing to study observations is a significant oversight because observations underpin conclusions. Apparent unreliability or biasability at the level of conclusions may actually lie deeper, at the observational level, which would require different intervention. Of course, if research revealed only perfect reliability and minimal biasability at the level of conclusions, there may be little need to conduct similar research at the level of observations. But, as reviewed above, the limited data available suggests that forensic psychological assessments are less than perfectly reliable, and at least sometimes vulnerable to bias. Therefore, it remains critical to disentangle observations and conclusions in order to study them both.

In forensic science, the distinction between observations and conclusions is usually fairly clear. For example, a fingerprint expert will observe the minutia (distinct and well-defined individual characteristics; see Figure 1) in the friction ridge of the fingerprints. The expert then forms a conclusion as to whether they "match" based on whether the observed minutia in the fingerprints are "similar enough."[4]

Recognizing this distinction between observations and conclusions, the forensic science literature has revealed problems at the observation levels 1–4 within and between experts, and with regard to reliability and biasability. For example, fingerprint examiners observe different data in the evidence depending on irrelevant contextual information (Earwaker et al., 2015), and they observe different data in the same evidence, between and within themselves, even without irrelevant contextual information (see Dror et al., 2011). These types of studies measure different aspects of expert performance, quantify them, and reveal the phenomena that contribute to differences at the observational level. These phenomena may arise from human factors (e.g., training and methods, exposure to irrelevant information) and/or from the data itself (i.e., problems of unreliability and biasability are most apparent when the data is of low quality). But studies that explore these phenomena can inform training and best practices to minimize such differences (e.g., the Fingerprint Analyses Consistency Tester, which helps to quantify and calibrate expert observations; Dror et al., 2011).

In forensic psychology, the distinction between observations and conclusions can be more complicated than in forensic science. For example, assigning a clinical diagnosis or reaching an opinion about a defendant's adjudicative competence or legal sanity are certainly *conclusions,* based on observations from a clinical interview, record review, and other sources. But each final conclusion (e.g., is the defendant competent or incompetent?) rests on numerous intermediate conclusions (e.g., does the defendant factually understand the legal process and his charges? does he rationally understand these? can he assist counsel?), *which depend on ob-*

[4] This—whether two prints are "similar enough"—is a subjective determination, because most jurisdictions have no definition or criteria as to what constitutes "similar enough" (Dror & Cole, 2010). Having such subjective criteria may contribute to problems of biasability and unreliability. But similarly subjective criteria are common in forensic psychology. Forensic psychologists may observe a defendant's deficits or impairments, but the law provides no precise guidance as to when the deficits are "deficient enough" to conclude that the defendant is incompetent to stand trial.

servations (e.g., what information in records is relevant to his capacities? what symptoms did he display during interview?).

Likewise, assigning an overall score on a complex assessment instrument—such as Hare's (2003) Psychopathy Checklist-Revised (PCL-R)—is a conclusion, reached only after countless observations necessary to score each of the instrument's 20 items. But even scoring many of the individual PCL-R items can be viewed as a *conclusion*, based upon many observations and inferences during interview and record review to decide whether a particular criterion—such as superficial charm, pathological lying, shallow emotions, or lack of remorse—should be scored as 0 (*absent*), 1 (*partially present*), or 2 (*clearly present*).

On the other hand, some items on forensic assessment instruments are much more similar to observations than conclusions. For example, the Static-99R (Helmus, Thornton, Hanson, & Babchishin, 2012) includes items such as "age at release" and "any male victims [in the examinee's sexual offense history]" that require relatively straightforward observations from criminal records. Thus, it sometimes seems clear what is an observation and what is a conclusion; however, other times it is difficult to clearly delineate observations from conclusions.

Forensic psychology research has almost solely focused on expert conclusions and has for the most part neglected researching expert observations. Indeed, our review identified *no* studies of forensic mental health evaluation that specifically addressed Levels 1 to 4 of observations. The only observation-level data that the field appears to offer are certain item-level data from test instrument manuals or test reliability studies. Again, some items on risk measures like the Static-99R (Helmus et al., 2012) are essentially coded observations from criminal records; these tend to show high reliability (Phenix & Epperson, 2015; Phenix, Helmus, & Hanson, 2015). Likewise, item-level data in Hare's (2003) PCL-R manual reveals stronger reliability values for those few items that are more like *observations* from the criminal record (e.g., juvenile delinquency, revocation of conditional release; ICC_{A1} = .75–.80) than those items that are *conclusions* about behavior (e.g., impulsivity, glibness, callousness; ICC_{A1} = .23–.36; see Hare, 2003; Sturup et al., 2014). Generally, on forensic assessment instruments that require clinician inference, interrater reliability and predictive validity are both stronger for less subjective items that are more like observations, and weaker for those more subjective items that are more like conclusions (Rufino, Boccaccini, & Guy, 2011) Of course, even ostensibly simple observations may depend on how the data are collected and examined, and these "data collection" procedures (e.g., how a forensic evaluator asks a defendant questions, or which collateral sources a forensic evaluator seeks and prioritizes) may be vulnerable to biases. Furthermore, many of data sources that evaluators use to score these instruments (e.g., police, defendant, victim, or witness statements; prior evaluation reports) may themselves have been influenced by various biases.

Beyond item-level data from forensic assessment instruments, we know of no reliability data for the many observations that inform forensic psychological evaluations, particularly those requiring clinical expertise (e.g., observing symptoms of psychosis, observing indicators of past mental state in collateral records, observing evidence of malingering), nor any data on the biasing effects that irrelevant contextual information may have on such observations. A broad review of interrater reliability estimates across a variety of medical and psychological procedures suggests that what we would consider *observations*—that is, "circumscribed judgment tasks requiring relatively few bits of information—such as test scoring, object counts, or physical measurements (e.g., counting decayed teeth, measuring organ size on ultrasound)"—tend to show stronger reliability than what we could call *conclusions*—"complex tasks requiring the synthesis of multiple, higher inferences (e.g., job performance ratings, stroke classification by neurologists)"; see Meyer, Mihura, & Smith (2005, p. 310). Nevertheless, even reliability for simple observations or judgments was imperfect, suggesting that we should not take observation-level reliability for granted in forensic psychology.

Discussion

The forensic sciences have long focused solely on the objects of their inquiries (DNA, fingerprints, firearms, handwriting, etc.) while neglecting—if not totally ignoring—the critical role that the human experts play in forensic decision making. However, the past decade has seen a dramatic shift in the forensic sciences, now recognizing, researching, developing policies, and mandating changes to reduce variability in expert decision making.

Forensic psychology has also tended to focus on the object of their inquiries—human behavior vis-à-vis legal standards—with far less focus on the critical role that the actual forensic psychologists—as human expert examiners—play in forensic assessments. To be fair, forensic psychology has historically explored certain aspects of reliability (e.g., Poythress & Stock, 1980), particularly in the context of psychological assessment instruments, long before the recent reforms in the forensic sciences. Nevertheless, for a field so rooted in the study of human behavior, cognition, and psychology, there has been surprisingly little attention to the role of human experts and human decision making in forensic psychological assessment. Put bluntly, the field tends to value reliability and objectivity, but tends to consider these more as qualities to be studied and maximized in instruments, with less attention to studying and maximizing these among the human experts rendering forensic opinions.

The revolution in forensic science has brought about scrutiny and examination of what factors influence forensic science experts. Research in this area has demonstrated that variability in forensic science decision making arises from two distinct factors: basic *reliability* (i.e., repeatability, reproducibility, consistency in decision making) and *biasability* (being inappropriately influenced by task-irrelevant information) (Dror & Rosenthal, 2008). Research in this area has further disentangled different components in expert forensic science decision making between the *conclusions* (e.g., whether the fingerprints match; Ulery et al., 2012) versus the *observations* on which the conclusions are based (e.g., what characteristics are observed in the fingerprint; Dror et al., 2011). Finally, the performance and variability of forensic science experts has been examined and quantified *between-experts* (variability among experts, the intervariability performance; Dror & Hampikian, 2011) and *within-experts* (variability within a single expert, the intravariability performance; Dror & Rosenthal, 2008). Combining these elements yields an eight-level framework for expert decision making—the Hierarchy of Expert Performance (HEP; Dror, 2016, see Figure 2), which has helped organize and frame the existing research, shown gaps where further research is

needed, and identified specific problematic areas that require improved policies and practices.

Applying HEP to forensic psychology reveals a few areas of relative strength, but more areas in which basic research is sorely lacking. Regarding relative strengths, forensic psychology has offered some research regarding *reliability between experts' conclusions*. This research comprises many studies detailing reliability in scoring specific instruments (e.g., Otto et al., 1998; Rogers et al., 2003) and a few studies documenting the field reliability of common criminal forensic evaluations such as those addressing competence and sanity (Guarnera & Murrie, 2017). But the field lacks adequate data for many other types of forensic evaluations, as well as other types of expert conclusions (e.g., those regarding diagnosis or psychological injury) that are central to many forensic evaluations.

Forensic psychology also lacks data at the level of *observations* (in contrast to conclusions). Decades ago, the influential jurist David Bazelon (1982) noticed this weakness and warned psychologists,

> Behavioral scientists who appear in the public arena all too often focus on little more than making conclusory pronouncements. Either they omit any real discussion of underlying observations and methods of inference, or they drown such discussion in a sea of jargon . . .
>
> What the public needs most from any expert, including the psychologist, is a wealth of intermediate observations and conceptual insights that are adequately explained. (p. 116)

Though Bazelon's primary concern was that individual experts disclose the methods, limits, and values underlying their work, his critique remains apt for research has well. We have some data to shed light on evaluator conclusions, but almost none to shed light on the "intermediate observations" on which conclusions should rest.

Consider a practical example: is the reliability in sanity conclusions modest because evaluators disagree in how they make a final inference regarding mental state at the time of offense? Or because evaluators disagree even earlier in the process, by reviewing different sources of information and observing different data in those sources? No available research sheds light on such critical questions, though it is plausible to imagine studies that could do so. For example, studies can remove the observational components from the forensic evaluation: in such studies, the same observations will be provided to the examiners (rather than the data and information which they use to make the observations), hence ensuring that they all start with the same observations; any differences can then be attributed to their inferences rather than to differences in their observations. In contrast to examining such differences in inferences, one can study the observations per se by providing examiners with identical records or videos of interviews, and comparing the evaluators' observations of the data. Furthermore, one can research the interactions between the observations and conclusions, for example, studying whether changes in observations drive changes in conclusions (as they should) versus situations in which expectations about conclusions (e.g., prompted by research designs that provide irrelevant information) influence what data is observed (a biasing effect of working backward, or circular reasoning, which the LSU approach was designed to minimize in forensic science, Dror et al., 2015).

This type of research has the potential to inform tools and resources for evaluators (such as checklists to guide procedures; see Gawande, 2010) or even policies governing evaluations (e.g., mandating that evaluators receive and consider certain uniform information from records). Efforts to study and enhance observation-level reliability could fit practically into many forensic training programs, whether for early stage trainees like graduate students or for practicing professionals participating in continuing education. Indeed, more training emphasis on reliability at the level of observations will likely improve reliability at the level of conclusions. In short, data at the level of observations is conspicuously absent from forensic psychology research, but addressing this gap in the research may be relatively simple.

Whether at the level of observations or conclusions, the field seems to offer no data on reliability *within* experts. This fundamental, foundational form of reliability (see Kraemer et al., 2012) must be examined and quantified rather than presumed. Again, we acknowledge that such studies are challenging, but they are not impossible. Researchers might use case materials (e.g., psychological testing results, collateral records) rather than actual defendants, and incorporate reliability research into training or continuing education programs (see, e.g., Blais, Forth, & Hare, 2017 for an example of incorporating reliability research—albeit *between* experts—into training). Should studies reveal poor within-expert reliability (as have some studies in the forensic sciences), results may help inform more rigorous procedures (checklists, protocols, etc.) for forensic evaluations and early training.

The final domain in which the HEP reveals clear gaps in forensic psychology research is biasability. The available research on bias in forensic evaluation has addressed adversarial allegiance (Murrie & Boccaccini, 2015), a clear threat in adversarial justice systems such as those in the U.S., U.K., and many other countries.[5] But even this research body is small, and limited only to certain types of evaluations (particularly sex offender risk assessment).

Although adversarial allegiance may have received more research attention than other forms of bias, this is certainly not the only threat to objective evaluations. Biases related to race, sex, sexual orientation, age, disability, and religion have, to our knowledge, *never* been explored among forensic evaluators. This research gap is striking, considering that these potential biases are such a popular foci of other types of psychological research, and even forensic psychology research addressing jury-decision making (e.g., Sommers & Norton, 2008). Other potential biases—for example, basic base-rate expectation biases, or biases related to crime details or criminal stereotypes—are currently understudied. Indeed, the field is increasingly attuned to many ways in which forensic evaluators may be vulnerable to bias (see Neal & Grisso, 2014; Zapf & Dror, 2017 for reviews), but empirical research on these biases lags behind.

[5] We do not claim that alternative (nonadversarial) justice systems, such as the inquisitorial system, are better overall. Though less vulnerable to adversarial allegiance, they may be more vulnerable to other biases or they may sacrifice strengths inherent in the adversarial system. Comparative research may reveal relative strengths and weaknesses of each, or interventions that the adversarial legal system can learn from other systems (such as expert "hot-tubbing;" Edmond, 2009).

Policy Implications

Our goal in this review has been primarily to provide a conceptual and practical framework for understanding and studying the performance of forensic evaluators. Without such a framework, it is easy to overlook the gaps and limitations in the available literature, which then leaves us more likely to pursue policies and practice that overlook (or even exacerbate) underlying problems. Indeed, our review of forensic psychology using the HEP reveals that there are many gaps in our knowledge base that prevent us from prescribing specific remediation in many areas.

Thus, the first priority should be performing research (some of which we have recommended throughout this review) that helps us better understand problems of unreliability and biasability within and between experts, at the levels of observations and conclusions. That said, some of this recommended research could *also* serve as pilot testing for certain policy or practice interventions. Likewise, certain policy or intervention studies might identify the nature of underlying problems, much like intervention studies with certain pharmaceuticals or medical procedures can shed light on the mechanisms underlying a disease. Therefore, we provide general suggestions for potential studies of interventions or policies that may help us better understand underlying unreliability and bias.

First, studies of existing policy arrangements are valuable. For example, the few jurisdictions that assign more than one evaluator per case (see Gowensmith, Pinals, & Karas, 2015) allow for studies of field reliability under various conditions. Take for example the studies of Hawaii's three-evaluator system (Gowensmith et al., 2012, 2013; 2017), which revealed that real-world reliability may be modest even in arrangements that have minimized the biasability attributable to adversarial allegiance. To take another example, some hospitals and at least one state (i.e., Virginia) have developed oversight policies that monitor conclusions from each of their evaluators on every assigned evaluation; these may allow for naturalistic reliability-type studies examining *patterns* of findings across evaluators.

Second, addressing bias in particular, forensic psychology may benefit from considering the interventions and policies recently emerging from the forensic sciences. After documenting the powerful influence of irrelevant contextual information (Kassin, Dror, & Kukucka, 2013), and identifying specific mechanisms such as *bias cascade* versus *bias snowball* (Dror, Morgan, Rando, & Nakhaeizadeh, 2017), the field is working to distinguish task-relevant from task-irrelevant information (National Commission on Forensic Science, 2015), developing policies and procedures to shield analysts from the latter. Indeed, these efforts have resulted in well-developed strategies for laboratories and agencies to process cases and evidence in ways that minimize analyst exposure to potentially biasing information; these include the use of case managers (Dror, 2013) and Linear Sequential Unmasking (LSU; Dror et al., 2015).

Such research addressing bias and best practices has been adopted in forensic science policy in the United States (NCFS, 2015) and in the United Kingdom (Forensic Science Regulator, 2015). Again, we acknowledge that distinguishing task-relevant from task-irrelevant information is more challenging in forensic psychology, but it is no less important. Furthermore, the policies do not only address issues of what is task irrelevant, but also 'when' task relevant information should be provided, that is, the best sequence and order that relevant information should be provided to examiners to minimize bias (such as circular and backward reasoning, see, e.g., the LSU policy). Trying to apply such de-biasing policies and procedures from the forensic sciences to forensic psychology is an important and promising first step. Some of the policies may be easily applied, others may require modifications and adaptations, whereas others may not be applicable. In any case, such an attempt is worthwhile and will help in creating policies and research strategy to better understand biasability, and ways to reduce bias.

Regarding the adversarial allegiance bias, the most commonly recommended policy is to explore court-appointed experts. Though intuitively appealing, court-appointed experts bring new challenges and dilemmas (Mnookin, 2008; Murrie & Boccaccini, 2015) that no research has yet addressed. However, researchers could explore a few domains, particularly child custody litigation, in which informal policies have begun to shift toward court-appointed, or jointly appointed experts over adversarial, opposing experts. Researchers might explore other appealing, but untested legal reform proposals such as "blinding" experts to the side retaining their services (Robertson & Kesselheim, 2016; Robertson & Yokum, 2012).[6] Again, we do not necessarily recommend any of these as broad policy reforms because there are not yet sufficient data to support them, but there is clearly value in exploring some of these in order to shed light on unreliability and biasability, as well as their potential solutions.

Conclusion

Forensic psychology offers well-developed procedures for the expert assessment of criminal defendants and civil litigants. But the field has offered much less data exploring the decision making of the forensic experts themselves. Recent reforms in the forensic sciences underscore the need to carefully study forensic experts, and Dror's (2016) HEP conceptualizes and defines the aspects involved in expert decision making, thus helping to frame the existing research and identify gaps. Forensic psychology can learn from these insights and use HEP to benefit and enhance forensic psychology decision making.

[6] Of course, not all potential interventions would reduce allegiance effects. Neal and Grisso (2014) proposed a "thought experiment", in which adversarial experts simply present the most compelling case they can for the side that retained them, with no pretense of objectivity.

References

Aboraya, A., Rankin, E., France, C., El-Missiry, A., & John, C. (2006). The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry, 3,* 41–50.

American Educational Research Association, American Psychological Association, and National Council on Measurement Education. (2014). *The Standards for Educational and Psychological Testing.* Washington, DC: AERA Publications.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073. http://dx.doi.org/10.1037//0003-066X.57.12.1060

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *American Psychologist, 68,* 7–19. http://dx.doi.org/10.1037/a0029889

Bazelon, D. L. (1982). Veils, values, and social responsibility. *American Psychologist, 37,* 115–121. http://dx.doi.org/10.1037/0003-066X.37.2.115

Blais, J., Forth, A. E., & Hare, R. D. (2017). Examining the interrater reliability of the Hare Psychopathy Checklist-Revised across a large sample of trained raters. *Psychological Assessment, 29,* 762–775. http://dx.doi.org/10.1037/pas0000455

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A., & Jeglic, E. (2012). Implications of Static-99 field reliability findings for score use and interpretation. *Criminal Justice and Behavior, 39,* 42–58. http://dx.doi.org/10.1177/0093854811427131

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others?: Findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law, 14,* 262–283. http://dx.doi.org/10.1037/a0014523

Coble, M. (2015). *Interpretation errors detected in a NIST interlaboratory study on DNA mixture interpretation in the U.S. (MIX13).* Presentation at the International Symposium on Forensic Science Error Management: Detection, Measurement, and Mitigation, Washington, DC.

Douglas, K. S., & Ogloff, J. R. P. (2003). The impact of confidence on the accuracy of structured professional and actuarial violence risk judgments in a sample of forensic psychiatric patients. *Law and Human Behavior, 27,* 573–587. http://dx.doi.org/10.1023/B:LAHU.0000004887.50905.f7

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11,* 347–383. http://dx.doi.org/10.1037/1076-8971.11.3.347

Dror, I. E. (2013). Practical solutions to cognitive and human factor challenges in forensic science. *Forensic Science Policy & Management: An International Journal, 4*(3–4), 105–113. http://dx.doi.org/10.1080/19409044.2014.901437

Dror, I. E. (2016). A hierarchy of expert performance. *Journal of Applied Research in Memory & Cognition, 5,* 121–127. http://dx.doi.org/10.1016/j.jarmac.2016.03.001

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. *Forensic Science International, 208*(1–3), 10–17. http://dx.doi.org/10.1016/j.forsciint.2010.10.013

Dror, I. E., & Cole, S. A. (2010). The vision in "blind" justice: Expert perception, judgment, and visual cognition in forensic pattern recognition. *Psychonomic Bulletin & Review, 17,* 161–167. http://dx.doi.org/10.3758/PBR.17.2.161

Dror, I. E., & Hampikian, G. (2011). Subjectivity and bias in forensic DNA mixture interpretation. *Science & Justice: Journal of the Forensic Science Society, 51,* 204–208. http://dx.doi.org/10.1016/j.scijus.2011.08.004

Dror, I. E., Morgan, R. M., Rando, C., & Nakhaeizadeh, S. (2017). The bias snowball and the bias cascade effects: Two distinct biases that may impact forensic decision making. *Journal of Forensic Sciences, 62,* 832–833. http://dx.doi.org/10.1111/1556-4029.13496

Dror, I., & Rosenthal, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *Journal of Forensic Sciences, 53,* 900–903. http://dx.doi.org/10.1111/j.1556-4029.2008.00762.x

Dror, I. E., Thompson, W. C., Meissner, C. A., Kornfield, I., Krane, D., Saks, M., & Risinger, M. (2015). Context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. *Journal of Forensic Sciences, 60,* 1111–1112. http://dx.doi.org/10.1111/1556-4029.12805

Earwaker, H., Morgan, R. M., Harris, A. J. L., & Hall, L. J. (2015). Fingermark submission decision-making within a UK fingerprint laboratory: Do experts get the marks that they need? *Science & Justice: Journal of the Forensic Science Society, 55,* 239–247. http://dx.doi.org/10.1016/j.scijus.2015.01.007

Edens, J. F., & Boccaccini, M. T. (2017). Taking forensic mental health assessment "out of the lab" and into "the real world": Introduction to the special issue on the field utility of forensic assessment instruments and procedures. *Psychological Assessment, 29,* 599–610. http://dx.doi.org/10.1037/pas0000475

Edmond, G. (2009). Merton and the hot tub: Scientific conventions and expert evidence in Australian civil procedure. *Law and Contemporary Problems, 72,* 159–189. Retrieved from http://www.jstor.org/stable/40647170

Forensic Science Regulator. (2015). *Cognitive bias effects relevant to forensic science examinations: Guidance.* Birmingham: The Forensic Science Regulator. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/510147/217_FSR-G-217_Cognitive_bias_appendix.pdf

Foster, W. L. (1897). Expert testimony, prevalent complaints and proposed remedies. *Harvard Law Review, 11,* 169–186. http://dx.doi.org/10.2307/1321970

Garrett, B. L., & Neufeld, P. J. (2009). Invalid forensic science testimony and wrongful convictions. *Virginia Law Review, 95,* 1–97.

Gawande, A. (2010). Checklists for success inside the OR and beyond: An interview with Atul Gawanda, MD, FACS. Interview by Tony Peregrin. *Bulletin of the American College of Surgeons, 95,* 24–27.

Gowensmith, W. N., Murrie, D. C., & Boccaccini, M. T. (2012). Field reliability of competence to stand trial opinions: How often do evaluators agree, and what do judges decide when evaluators disagree? *Law and Human Behavior, 36,* 130–139. http://dx.doi.org/10.1037/h0093958

Gowensmith, W. N., Murrie, D. C., & Boccaccini, M. T. (2013). How reliable are forensic evaluations of legal sanity? *Law and Human Behavior, 37,* 98–106. http://dx.doi.org/10.1037/lhb0000001

Gowensmith, W. N., Murrie, D. C., Boccaccini, M. T., & McNichols, B. J. (2017). Field reliability influences field validity: Risk assessments of individuals found not guilty by reason of insanity. *Psychological Assessment, 29,* 786–794. http://dx.doi.org/10.1037/pas0000376

Gowensmith, W. N., Pinals, D. A., & Karas, A. C. (2015). States' standards for training and certifying evaluators of competency to stand trial. *Journal of Forensic Psychology Practice, 15,* 295–317. http://dx.doi.org/10.1080/15228932.2015.1046798

Gowensmith, W. N., Sessarego, S. N., McKee, M. K., Horkott, S., MacLean, N., & McCallum, K. E. (2017). Diagnostic field reliability in forensic mental health evaluations. *Psychological Assessment, 29,* 692–700. http://dx.doi.org/10.1037/pas0000425

Guarnera, L., & Murrie, D. C. (2017). Field reliability of adjudicative competence and legal sanity opinions: A systematic review and meta-analysis. *Psychological Assessment, 29,* 795–818. http://dx.doi.org/10.1037/pas0000388

Guarnera, L., Murrie, D. C., & Boccaccini, M. T. (2017). Why do forensic experts disagree? Sources of unreliability and bias in forensic psychology evaluations. *Translational Issues in Psychological Science, 3,* 143–152.

Hand, L. (1901). Historical and practical considerations regarding expert testimony. *Harvard Law Review, 15,* 40–58. http://dx.doi.org/10.2307/1322532

Hare, R. D. (2003). *The Hare Psychopathy Checklist–Revised* (2nd ed.). Toronto, Ontario: Multi-Health Systems.

Hawaii Revised Statutes, Vol. 14, §704–111 (2014).

Heilbrun, K., & Brooks, S. (2010). Forensic psychology and forensic science: A proposed agenda for the next decade. *Psychology, Public Policy, and Law, 16,* 219–253. http://dx.doi.org/10.1037/a0019138

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse, 24,* 64–101. http://dx.doi.org/10.1177/1079063211409951

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied*

Research in Memory & Cognition, 2, 42–52. http://dx.doi.org/10.1016/j.jarmac.2013.01.001

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). *DSM–5*: How reliable is reliable enough? *The American Journal of Psychiatry, 169,* 13–15. http://dx.doi.org/10.1176/appi.ajp.2011.11010050

Melton, G. B., Petrila, J., Poythress, N. G., & Slobogin, C. (2007). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers.* New York, NY: Guilford Press.

Meyer, G. J., Mihura, J. L., & Smith, B. L. (2005). The interclinician reliability of Rorschach interpretation in four data sets. *Journal of Personality Assessment, 84,* 296–314. http://dx.doi.org/10.1207/s15327752jpa8403_09

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior, 29,* 621–637. http://dx.doi.org/10.1007/s10979-005-8122-9

Mnookin, J. (2008). Expert evidence, partisanship, and epistemic confidence. *Brooklyn Law Review, 73,* 587–611.

Monahan, J., Heilbrun, K., Silver, E., Nabors, E., Bone, J., & Slovic, P. (2002). Communicating violence risk: Frequency formats, vivid outcomes, and forensic settings. *The International Journal of Forensic Mental Health, 1,* 121–126. http://dx.doi.org/10.1080/14999013.2002.10471167

Murrie, D. C., & Boccaccini, M. T. (2015). Adversarial allegiance among expert witnesses. *Annual Review of Law and Social Science, 11,* 37–55. http://dx.doi.org/10.1146/annurev-lawsocsci-120814-121714

Murrie, D. C., Boccaccini, M. T., Guarnera, L. A., & Rufino, K. A. (2013). Are forensic experts biased by the side that retained them? *Psychological Science, 24,* 1889–1897. http://dx.doi.org/10.1177/0956797613481812

Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior, 32,* 352–362. http://dx.doi.org/10.1007/s10979-007-9097-5

Murrie, D. C., Boccaccini, M. T., Turner, D. B., Meeks, M., Woods, C., & Tussey, C. (2009). Rater (dis)agreement on risk assessment measures in sexually violent predator proceedings: Evidence of adversarial allegiance in forensic evaluation? *Psychology, Public Policy, and Law, 15,* 19–53. http://dx.doi.org/10.1037/a0014897

Murrie, D. C., Boccaccini, M. T., Zapf, P. A., Warren, J. I., & Henderson, C. E. (2008). Clinician variation in findings of competence to stand trial. *Psychology, Public Policy, and Law, 14,* 177–193. http://dx.doi.org/10.1037/a0013578

Murrie, D. C., & Warren, J. I. (2005). Clinician variation in rates of legal sanity opinions: Implications for self-monitoring. *Professional Psychology: Research and Practice, 36,* 519–524. http://dx.doi.org/10.1037/0735-7028.36.5.519

National Commission on Forensic Science. (2015). *Ensuring that forensic analysis is based upon task relevant information.* National Institute of Standards and Technology. Retrieved from https://www.justice.gov/ncfs/file/818196/download

National Institute of Standards and Technology. (2016). *New NIST Center of Excellence to Improve Statistical Analysis of Forensic Evidence.* Retrieved May 30, 2017 from https://www.nist.gov/news-events/news/2015/05/new-nist-center-excellence-improve-statistical-analysis-forensic-evidence

National Research Council, Committee on Identifying the Needs of the Forensic Science Community. (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: The National Academies Press. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf

Neal, T. M. S., & Grisso, T. (2014). The cognitive underpinnings of bias in forensic mental health evaluations. *Psychology, Public Policy, and Law, 20,* 200–211. http://dx.doi.org/10.1037/a0035824

Otto, R. K., Poythress, N. G., Nicholson, R. A., Edens, J. F., Monahan, J., Bonnie, R. J., . . . Eisenberg, M. (1998). Psychometric properties of the MacArthur Competence Assessment Tool–Criminal Adjudication. *Psychological Assessment, 10,* 435–443. http://dx.doi.org/10.1037/1040-3590.10.4.435

Packer, I. K. (2009). *Evaluation of criminal responsibility.* New York, NY: New York: Oxford University Press. http://dx.doi.org/10.1093/med:psych/9780195324853.001.0001

Phenix, A., & Epperson, D. L. (2015). Overview of the development, reliability, validity, scoring, and uses of the Static-99, Static-99R, Static-2002, and Static-2002R. In A. Phenix & H. M. Hoberman (Eds.), *Sexual offending: Predisposing conditions, assessments, and management* (pp. 437–455). New York, NY: Springer.

Phenix, A., Helmus, H., & Hanson, R. K. (2015). Static-99R and Static-2002R evaluators' workbook [Unpublished manual]. Retrieved from www.static99.org

Poythress, N. G., & Stock, H. V. (1980). Competency to stand trial: A historical review and some new data. *The Journal of Psychiatry & Law, 8,* 131–146.

President's Council of Advisors on Science and Technology. (2016). *Report to the President: Forensic science in the criminal courts: Ensuring scientific validity of feature-comparison methods.* Washington, DC: Executive Office of the President of the United States. Retrieved from https://www.theiai.org/president/201609_PCAST_Forensic_Science_Report_FINAL.pdf

Robertson, C. T., & Kesselheim, A. S. (2016). *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law.* San Diego, CA: Elsevier Inc.

Robertson, C. T., & Yokum, D. V. (2012). The effect of blinded experts on juror verdicts. *Journal of Empirical Legal Studies, 9,* 765–794. http://dx.doi.org/10.1111/j.1740-1461.2012.01273.x

Rogers, R., Jackson, R. L., Sewell, K. W., Tillbrook, C. E., & Martin, M. A. (2003). Assessing dimensions of competency to stand trial: Construct validation of the ECST-R. *Assessment, 10,* 344–351. http://dx.doi.org/10.1177/1073191103259007

Rufino, K. A., Boccaccini, M. T., & Guy, L. S. (2011). Scoring subjectivity and item performance on measures used to assess violence risk: The PCL-R and HCR-20 as exemplars. *Assessment, 18,* 453–463. http://dx.doi.org/10.1177/1073191110378482

Scurich, N., Monahan, J., & John, R. S. (2012). Innumeracy and unpacking: Bridging the nomothetic/idiographic divide in violence risk assessment. *Law and Human Behavior, 36,* 548–554. http://dx.doi.org/10.1037/h0093994

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24,* 271–296. http://dx.doi.org/10.1023/A:1005595519944

Smalarz, L., Madon, S., Yang, Y., Guyll, M., & Buck, S. (2016). The perfect match: Do criminal stereotypes bias forensic evidence analysis? *Law and Human Behavior, 40,* 420–429. http://dx.doi.org/10.1037/lhb0000190

Sommers, S. R., & Norton, M. I. (2008). Race and jury selection: Psychological perspectives on the peremptory challenge debate. *American Psychologist, 63,* 527–539. http://dx.doi.org/10.1037/0003-066X.63.6.527

Spitzer, R. L., & Fleiss, J. L. (1974). A re-analysis of the reliability of psychiatric diagnosis. *The British Journal of Psychiatry, 125,* 341–347. http://dx.doi.org/10.1192/bjp.125.4.341

Sturup, J., Edens, J. F., Sörman, K., Karlberg, D., Fredriksson, B., & Kristiansson, M. (2014). Field reliability of the Psychopathy Checklist-

Revised among life sentenced prisoners in Sweden. *Law and Human Behavior, 38,* 315–324. http://dx.doi.org/10.1037/lhb0000063

Thomas, C. M., Bertram, E., & Johnson, D. (2009). The SBAR communication technique: Teaching nursing students professional communication skills. *Nurse Educator, 34,* 176–180. http://dx.doi.org/10.1097/NNE.0b013e3181aaba54

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE, 7*(3), e32800. http://dx.doi.org/10.1371/journal.pone.0032800

Wacogne, I., & Diwakar, V. (2010). Handover and note-keeping: The SBAR approach. *Clinical Risk, 16,* 173–175. http://dx.doi.org/10.1258/cr.2010.010043

Weed, L. L. (1970). *Medical records, medical evaluation, and patient care: The problem-oriented medical record as a basic tool.* Cleveland, OH: Press of Case Western Reserve University.

Wigmore, J. H. (1923). *A treatise on the Anglo-American system of evidence in trials at common law: Including the statutes and judicial decisions of all jurisdictions of the United States and Canada.* Boston, MA: Little, Brown.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The comprehensive system for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00658.x

Zapf, P. A., & Dror, I. E. (2017). Understanding and mitigating bias in forensic evaluation: Lessons from forensic science. *International Journal of Forensic Mental Health.*

Zapf, P. A., & Roesch, R. (2009). *Evaluation of competency to stand trial.* New York, NY: Oxford University Press.

Received March 20, 2017
Revision received June 22, 2017
Accepted June 23, 2017 ■

Vocational Psychology Assessment

Positive Human Characteristics Leading to Positive Work Outcomes

Christine Robitschek and Matthew W. Ashton

ork can play many roles in a person's life: It can be a way of earning money for survival or to support a chosen lifestyle, a pathway on which a person progresses (e.g., earns promotions and recognition), or a mechanism by which one expresses purpose in life and self-concept (Super, 1963; Wrzesniewski, McCauley, Rozin, & Schwartz, 1997). Work provides benefits to both the individual engaging in the work and to society (e.g., Gerstel & Gross, 1987), which reflects positive psychology's shared emphases on personal and societal well-being (Seligman & Csikszentmihalyi, 2000). If workers are to strive for positive outcomes for themselves and society, however, they must possess or develop positive human characteristics and behaviors. This chapter addresses these positive characteristics, behaviors, and outcomes.

Because of space limitations, this chapter does not cover the breadth of constructs or perspectives in the work domain. A few examples of relevant topics that the reader may find interesting but that are not covered are Savickas's (2000) taxonomy of human strengths, which is derived from vocational theory and can be applied across life domains; Wrzesniewski et al.'s (1997) assessment of work as job, career, or calling; and Sympson's (1999) operationalization of hope in the work domain. Given the expertise of the authors, this chapter focuses on the assessment of constructs found in the vocational psychology literature.

The following sections describe assessment of good career decision making, the role of work in our lives, adaptability within the career role, and areas for

http://dx.doi.org/10.1037/0000138-023
Positive Psychological Assessment: A Handbook of Models and Measures, Second Edition,
M. W. Gallagher and S. J. Lopez (Editors)

future assessment efforts. Although 12 assessments are described, many more remain unmentioned. The reader is directed to Kapes, Mastie, and Whitfield (1994), Seligman (1994), and Levinson, Ohlers, Caswell, and Kiewra (1998) for descriptions of many other measures.

HOW PEOPLE MAKE "GOOD" CAREER DECISIONS

The bulk of vocational psychology literature deals with the way in which human beings go about making decisions within and about their careers. Phillips and Jome (2005) summarized the vocational literature about career decision making, noting that the "best" career choices may be defined by either (a) an individual's selecting the "best" option for him or her, or (b) an individual's engaging in the "best" decision-making process regardless of what alternative is selected. We discuss the different constructs and processes that have been most prominently connected to making "good" career decisions and the assessment instruments that measure them.

Career Exploration

Most current theories of career development and choice highlight the importance of *career exploration*, defined as behavior that increases individuals' understanding of themselves or their environment with the aim of choosing or progressing within an occupation (Jordaan, 1963). Exploratory behaviors are beneficial at predictable developmental stages (i.e., adolescence and emerging adulthood) characterized by work-related experimentation (Super, Savickas, & Super, 1996) and during career transitions in which the behaviors assist the person with important decisions. This traditional view of career exploration is similar to *exploration in breadth*, which is gathering information about a variety of options to be used in making decisions (Luyckx, Goossens, Soenens, & Beyers, 2006). *Exploration in depth* is also important. This involves exploring one's current work commitments (Luyckx et al., 2006), which is an important behavior as people reevaluate their commitments, for example, as a job or work environment changes. Given these perspectives, assessment of career exploration should address exploration of the self and the work environment as well as exploration in breadth and exploration in depth.

Career Exploration Survey

The Career Exploration Survey (CES; Stumpf, Colarelli, & Hartman, 1983) is a 59-item instrument that is administered and scored by the researcher or practitioner. Test takers answer items in the context of the 3 months before taking the CES, responding to each item on a 5-point scale, with anchors that vary to match item content. For example, 1 = *little* or *not satisfied* and 5 = *a great deal* or *very satisfied*. Results yield scores on 16 dimensions of career exploration. Several dimensions are aspects of the exploration process: environmental exploration, self-exploration, number of occupations considered,

intended-systematic exploration, frequency (of exploratory behavior), amount of information, and focus. Three dimensions are aspects of reactions to exploration: satisfaction with information, explorational stress, and decisional stress. Six dimensions are aspects of beliefs: employment outlook, certainty of career exploration outcome, external search instrumentality, internal search instrumentality, method instrumentality, and importance of obtaining preferred position. The CES provides a multidimensional perspective on exploration. (See Stumpf et al., 1983, for psychometric information.)

Vocational Identity Status Assessment

The Vocational Identity Status Assessment (VISA; Porfeli, Lee, Vondracek, & Weigold, 2011) is a 30-item measure of work-related identity status. The VISA assesses two broad dimensions of identity formation: career exploration and career commitment. Within career exploration, there are two subscales: In-Breadth Career Exploration and In-Depth Career Exploration. Each subscale comprises five items with response options ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Subscale scores are calculated by averaging item scores for that subscale. The measure and psychometric information are available in Porfeli et al. (2011).

Vocational Interests

Many current vocational theories emphasize the importance of vocational interests as a foundation for making career decisions. Interests can be an indicator of a person's vocational strengths—that is, areas in which the person is likely to be motivated to learn and perform at a high level. Holland's (1959) vocational theory is perhaps the most widely used theory for categorizing vocational interests. Holland posited a hexagonal model of interests, which he viewed as personality types. These six types (with examples of typical interests) are (a) realistic (e.g., mechanics, agriculture, and sports); (b) investigative (e.g., science and scholarly pursuits); (c) artistic (e.g., visual or culinary arts, creative writing, and drama); (d) social (e.g., teaching, counseling, and other helping professions); (e) enterprising (e.g., selling products, services, or ideas); and (f) conventional (e.g., typing, filing, and accounting). A person's profile of interests can be expressed by scores on relatively independent scales measuring these six types. Profiles, arranged into what are known as Holland codes, typically consist of the three highest scores among these six types, although some people can best be described in fewer or greater numbers of types. In this section, we present two widely used measures of vocational interests based on Holland's model of interests.

Self-Directed Search

The Self-Directed Search, fifth edition (SDS; Holland & Messer, 2013), is administered (25–35 minutes) and scored (10 minutes) by the test taker. It is administered via paper and pencil (available from Psychological Assessment

Resources) or online at www.self-directed-search.com. Results yield Holland codes for "activities" (things the test taker would like to do), "competencies" (things the test taker already can do well), "occupations" (things in which the test taker has interest or finds appealing), and "self estimates" (self-ratings of abilities compared with other people). The test taker calculates a composite Holland code that includes all of these areas. There are several forms of the SDS: (a) Form R is the most commonly used form and is appropriate for ages 11 to 70, (b) Form E is written at a fourth-grade level for people with limited reading skills, and (c) Career Explorer is for junior high school and middle school students. Other forms are available in several languages. The SDS is used in conjunction with the *You and Your Career* booklet, which provides information about Holland's hexagonal model and assistance with career exploration; *Occupations Finder*, a booklet with a wide variety of occupations, listed by Holland code, as a means for test takers to compare their codes with the codes of occupations; and the *Educational Opportunities Finder, Veterans and Military Occupations Finder*, and *Leisure Finder*, which are used in similar ways. The online version of the SDS yields a Client Interpretive Report. Reliability and validity information is available in Holland, Fritzche, and Powell (1994).

Strong Interest Inventory

The Strong Interest Inventory, revised edition (SII; Consulting Psychologists Press, 2012), is a 291-item instrument that is administered by the researcher or practitioner and scored by the publisher, CPP. Test takers use a 5-point scale ranging from *strongly like* to *strongly dislike* to rate items in the five areas of occupations, school subjects, activities, leisure activities, and types of people. Test takers also mark the extent to which additional items identify their characteristics with response options on a 5-point scale ranging from *strongly like me* to *strongly unlike me*. Three sets of scores, related to Holland types, are provided in the results. General Occupational Themes are composite Holland codes. Basic Interest scales are subscales of the Holland codes. Occupational scales compare the test taker's profile with the profiles of people who are successfully employed in specific occupations.

The SII also yields scores on five bipolar Personal Style scales, which describe aspects of how the test taker prefers to interact with the world around him or her. The scales are work style, learning environment, leadership style, risk taking/adventure, and team orientation. Readers are directed to Donnay, Thompson, Morris, and Schaubhut (2004) and Herk and Thompson (2012) for reliability and validity information for the revised SII.

These two measures of vocational interests have different strengths. The SDS is useful in examining discrepancies among an individual's Holland codes as indicated by the different areas measured by the SDS. Areas include things I would like to do, things I already do well, things I find appealing, and self-ratings of my abilities compared to other people. The SII is particularly useful because it divides the test taker's Holland codes into the basic interest scales, which can help tease apart unexpected results. Also, the SII gives test

takers direct comparisons of their profiles with people in a wide range of occupations. In contrast, *Occupations Finder* of the SDS relates clients' Holland codes with the codes of occupations.

Work Needs and Values

Individuals vary in the things they need or want from work. One way to conceptualize these needs is in terms of basic psychological needs (e.g., needs identified in self-determination theory; Ryan & Deci, 2000) that may be satisfied through work settings and experiences. Ryan and Deci (2000) posited that the basic psychological needs of autonomy, competence, and relatedness are universal and inherent in human beings. When these basic needs are met in the workplace, people experience greater engagement in work and job satisfaction, more affective commitment to their work, and protection against burnout and job strain (see Van den Broeck, Ferris, Chang, & Rosen, 2016, for a review).

A second way of conceptualizing needs is in terms of *work values*, which are beliefs about the qualities of life that are considered important and desirable specifically in one's vocational pursuits (Hartung, 2009). Super (1990) viewed work values as acquired adaptations transmitted through proximal and distal cultural influences. When work environments are congruent with our work values, outcomes are similar to when our basic psychological needs are met. See, for example, greater job satisfaction (e.g., Feather & Rauter, 2004) and job commitment (e.g., Rounds, 1990).

Given these multiple conceptualizations of work needs and values, various measures serve different purposes in research and practice.

Work-Related Basic Need Satisfaction

The Work-Related Basic Need Satisfaction Scale (W-BNS; Van den Broeck, Vansteenkiste, De Witte, Soenens, & Lens, 2010) assesses the extent to which the needs for competence, autonomy, and relatedness are met in one's current work situation. The W-BNS has 16 items that the test taker rates on a Likert scale ranging from 1 (*totally disagree*) to 5 (*totally agree*). The three scale scores (each measuring one of the needs) are calculated by averaging the item scores for that scale. Psychometric information can be found in Van den Broeck and colleagues (2010). Although there is considerable research on self-determination theory and the importance of basic psychological need fulfillment across domains, only in recent years have we seen measures of work-related basic psychological needs, such as the W-BNS, with strong psychometric evidence. Also, we were unable to locate literature describing use of these measures in applied situations. Thus, although the W-BNS is useful in research, its utility in practice remains undetermined.

Super's Work Values Inventory—Revised

Super's Work Values Inventory—Revised (SWVI–r; Suen, 2015) is a 72-item instrument that is administered by the researcher or practitioner and scored

by the scale publisher (Kuder). Test takers indicate the level of importance for each item on a 5-point scale ranging from 1 (*Not important at all. Not a factor in my job selection.*) to 5 (*Crucial. I would not consider a job without it.*). Results yield scores on 12 scales (i.e., achievement, coworkers, creativity, income, independence, lifestyle, challenge, prestige, security, supervision, variety, and workplace). Reliability and validity information is summarized in Suen (2015). The SWVI–r has been used extensively in practice and research to aid in understanding how values can manifest in the work domain.

Work Values Inventory

The Work Values Inventory (WVI; Santa Cruz County Regional Occupational Program, n.d.) is a self-administered, self-scored, and self-interpreted assessment of work values. The WVI comprises four brief sections that assess core values in one's life (e.g., honesty and power), values related to work environments (e.g., flexible and high earnings), values related to coworker interactions (e.g., competition and diversity), and valuing types of work activities (e.g., creative and public contact). For each item with these four sections, respondents rate each item as *Always important, Sort of important,* or *Not important.* A fifth section of the WVI asks respondents to identify their top five values among the values they rated as *Always important.* Respondents then identify the section (e.g., core values, coworker interactions) each of these top values is from. Identifying the sections provides information regarding differential importance of broad domains of values for the respondent. A final section of the WVI asks respondents to "write a paragraph describing how you see your top 5 values being important in your work" (p. 3). This story helps the person to situate their work values specifically within their work experiences and goals. We were unable to locate any information regarding the psychometric properties of the WVI. However, in our experience, the WVI and similar assessments of work values are the assessments most commonly used in applied settings. These assessments are transparent to the test taker and are quickly administered and interpreted, without any need to submit the assessment elsewhere for scoring. This suggests a possible disconnect between research and practice in the assessment of work values and needs. Addressing this disconnect and conducting research on assessment of work values and needs that are connected to practice are warranted.

Job/Work Satisfaction

Assessing satisfaction with work in general or the current job a person holds continues to be a challenging task. A multitude of measures have been used in research. Yet it is not clear why researchers have opted to reinvent the wheel each time they want to measure satisfaction with work. Despite the plethora of measures with little evidence of reliability or validity, two measures that serve different purposes have seen some repeated use, the Minnesota Satisfaction Questionnaire (Weiss, Dawis, England, & Lofquist,

1967) and the Overall Job Satisfaction Scale (Judge, Locke, Durham, & Kluger, 1998).

Minnesota Satisfaction Questionnaire

The Minnesota Satisfaction Questionnaire (MSQ; Weiss et al., 1967) is administered and scored by the researcher or practitioner. The measure is available in the public domain from the publisher (Vocational Psychology Research, University of Minnesota–Minneapolis). Three forms are available: two versions of the 100-item long form (1967; 1977) and a 20-item short form (1977). The MSQ measures the degree to which test takers are satisfied with 20 aspects of their current job, such as recognition, security, advancement, and variety. Item responses on the 1967 long form are on a unidirectional 5-point scale ranging from *not satisfied* to *extremely satisfied*. The 1977 long and short forms use revised response options, a balanced 5-point scale ranging from *very satisfied* to *very dissatisfied*. Psychometric information is available in Weiss et al. (1967) and in the MSQ manual (Weiss et al., 1967).

Overall Job Satisfaction Scale

The Overall Job Satisfaction Scale (OJS; Judge et al., 1998) is administered and scored by the researcher or practitioner and is available in the public domain. The five items are based on a longer measure by Brayfield and Rothe (1951) and can be found in Judge et al. (1998). Item responses range from *strongly disagree* to *strongly agree*. Researchers and practitioners should use the OJS if they are interested in job satisfaction as a unitary construct. Users should consider the MSQ if they are interested in multidimensional work satisfaction.

ROLE SALIENCE AND BALANCE

Super (1980) put forth these constructs from a vocational perspective. *Role salience* refers to the absolute and relative importance of various life roles. *Role balance* refers to the extent to which a person is comfortable with the amount of time and energy put into each role in relation to other life roles. Our most salient life roles (referred to as *core* roles) are more critical to our life satisfaction than more peripheral life roles (Super et al., 1996). Therefore, if a core life role is demanding more time and energy than usual, we are able to sacrifice time and energy in peripheral life roles with limited effect on life satisfaction. Conversely, if peripheral roles demand more time and energy and core roles suffer as a result, this will have a negative impact on life satisfaction (Super et al., 1996).

Salience Inventory

The Salience Inventory (SI; Nevill & Super, 1986) is a 170-item instrument that is administered and scored by the researcher or practitioner, and it is available

free of charge to researchers and practitioners through www.vocopher.com. Item responses are on a 4-point scale from *never or rarely/little or none* to *almost always or always/a great deal.* Results yield scores for five life roles: student, worker, homemaker (including parenting and partner roles), leisurite, and citizen. Within each life role, three aspects of salience are tapped (yielding a total of 15 subscale scores: three aspects of salience for each of five life roles). These three aspects of salience are participation (i.e., what the test taker actually does in this life role), commitment (i.e., attitudinal and affective importance of the life role), and value expectations (i.e., the degree to which the life role is expected to fulfill the test taker's values and needs). Thus, the SI informs about not only which roles are most important but also the extent to which test takers actually are engaged in activities (participation) that are important (commitment) and meet their needs (values expectations). Reliability and validity information is available in Nevill and Super (1986).

Life-Career Rainbow

The Life-Career Rainbow (Super, 1980) is a qualitative way to assess role salience, among other constructs. Construction of the Rainbow can be completed by the individual being assessed after thorough instructions are given or by this individual in conjunction with the researcher or practitioner. The lifespan is represented by the length of the Rainbow, with the left and right ends representing birth and death, respectively. Each band of the Rainbow represents a different life role. The width of each band at any given point in the life span represents the salience of that life role at that point in time. For example, the "worker" band of the Rainbow likely would be empty for most people until sometime in the teenage years, at which point it might be fairly narrow (compared with other bands) if the worker role has minimal salience. In the adult years, the worker band might be wide, if, for example, the individual is employed full-time, outside the home, in a job that has meaning and purpose for the worker. This band is likely to narrow again or end completely after retirement depending on whether the person quits work altogether or continues to work in some part-time capacity after formally retiring. A cross section of the Rainbow at any point in the lifespan provides a picture of the life space (i.e., a comprehensive view of the multiple life roles a person plays at any one time).

Although the Life-Career Rainbow might be of limited utility to researchers, particularly those involved in quantitative research, it is very useful to practitioners and their clients. Similar to the SI, the Rainbow can help clients understand the relative importance of various life roles and how these roles might interact. In contrast to the SI, however, the Rainbow adds the life-span dimension, which allows people to explore how the importance of these roles, and even the presence or absence of each role, has changed over time. Furthermore, the Rainbow allows people to be planful about how they will structure their life space and balance their life roles in the future.

HOW INDIVIDUALS ADAPT AND CHANGE

The world of work is rapidly changing. A *typical* career path now involves multiple changes in job, employer, and often location over one's time in the workforce. Although many of these changes are instigated by the worker, many are not. Factors such as technological advances, abrupt economic recessions, outsourcing, and organizational mergers can result in vocational upheaval ranging from job restructuring to layoffs. Workers must be adaptive in their careers to cope successfully with these rapid changes (Murphy, Blustein, Bohlig, & Platt, 2010).

Career adaptability is an individual's readiness to handle both predictable and unexpected career changes and challenges across the lifespan (Super & Knasel, 1981). Here we present the most recent assessment of career adaptability, the Career Adapt-Abilities Scale (CAAS; Savickas & Porfeli, 2012).

The CAAS is a 24-item instrument available in Appendix 2 of Savickas and Porfeli (2012). The CAAS assesses four components of career adaptability: *Concern* about the future of one's career; taking *Control* and preparing for one's career; *Curiosity* about how one's career and self might be in the future; and *Confidence* in one's ability to achieve career goals. Response options range from 1 (*not strong*) to 5 (*strongest*). Scores are calculated by averaging item scores on each subscale. Initial psychometric information for the assessment is available in Savickas and Porfeli (2012). Additional psychometric information specific to each of 13 countries is available in multiple articles compiled in a special issue of the *Journal of Vocational Behavior* (Vol. 80, Issue 3).

MEASUREMENT ISSUES AND FUTURE DEVELOPMENTS

Considerable evidence indicates solid reliability and validity for the instruments discussed in this chapter. Readers are directed to the references mentioned throughout for test-specific measurement issues. In recent years, however, several themes have emerged questioning the breadth of utility for these measures. First, the majority of the psychometric evidence is for primarily White, non-Hispanic Americans. We need to devote considerable effort and resources to determine not only whether the assessments themselves are psychometrically sound for diverse populations but also whether the theoretical propositions underlying these assessments are culturally appropriate for people with diverse identities (Hardin, Robitschek, Flores, Navarro, & Ashton, 2014).

A second theme involves the changing world of work. Many of today's jobs did not exist 30 years ago (e.g., web designer or Transportation Security Administration baggage screener), when many of the assessments described here were first developed. We also have seen dramatic changes in economic globalization, movement further into a postindustrial economy, and tremendous advances in technology resulting in enormous reductions in the number of jobs for unskilled and semiskilled workers (DeBell, 2006). Yet most vocational

assessments have not kept pace with these dramatic changes in the world of work. Most vocational measures could benefit from modernization to reflect new jobs, new organizational structures, and new individual work patterns within the world of work.

Finally, future developments in vocational assessment need to address the distinction between "what is" and "what might be" in a person's work life. Unfortunately, this has changed little since the first edition of this book was published. As Krumboltz (e.g., Mitchell & Krumboltz, 1996) pointed out, we do a disservice to people if vocational assessment limits their choices to options to which they already have been exposed. Vocational assessment, particularly in the context of positive psychology, should open doors and increase the range of options that people perceive in the world of work. Current vocational assessment tools do an excellent job of assessing "what is." We now need to add to these tools to include "measures of the possible."

REFERENCES

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology, 35,* 307–311. http://dx.doi.org/10.1037/h0055617

Consulting Psychologists Press. (2012). *The Strong Interest Inventory* (Rev. ed.). Sunnyvale, CA: Author.

DeBell, C. (2006). What all applied psychologists should know about work. *Professional Psychology: Research and Practice, 37,* 325–333. http://dx.doi.org/10.1037/0735-7028.37.4.325

Donnay, D. A., Thompson, R. C., Morris, M. L., & Schaubhut, N. A. (2004). *Technical brief for the newly revised Strong Interest assessment: Content, reliability and validity.* Mountain View, CA: Consulting Psychologists Press.

Feather, N. T., & Rauter, K. A. (2004). Organizational citizenship behaviors in relation to job status, job insecurity, organizational commitment and identification, job satisfaction and work values. *Journal of Occupational and Organizational Psychology, 77*(1), 81–94. http://dx.doi.org/10.1348/096317904322915928

Gerstel, N., & Gross, H. E. (Eds.). (1987). *Families and work.* Philadelphia, PA: Temple University Press.

Hardin, E. E., Robitschek, C., Flores, L. Y., Navarro, R. L., & Ashton, M. W. (2014). The cultural lens approach to evaluating cultural validity of psychological theory. *American Psychologist, 69,* 656–668. http://dx.doi.org/10.1037/a0036532

Hartung, P. J. (2009, June). *Why work: The story of values in vocational psychology.* Invited address given at the Ninth Biennial Meeting of the Society for Vocational Psychology, St. Louis, MO.

Herk, N. A., & Thompson, R. C. (2012). *Strong Interest Inventory manual supplement: Occupational scales update 2012.* Mountain View, CA: Consulting Psychologists Press.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6,* 35–45. http://dx.doi.org/10.1037/h0040767

Holland, J. L., Fritzche, B. A., & Powell, A. B. (1994). *Technical manual for the Self-Directed Search.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L., & Messer, M. A. (2013). *Self-directed search* (5th ed.). Lutz, FL: Psychological Assessment Resources.

Jordaan, J. P. (1963). Exploratory behavior: The formation of self- and occupational concepts. In D. E. Super & Associates (Eds.), *Career development: Self-concept theory* (pp. 42–78). New York, NY: College Entrance Examination Board.

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology, 83,* 17–34. http://dx.doi.org/10.1037/0021-9010.83.1.17

Kapes, J. T., Mastie, M. M., & Whitfield, E. A. (Eds.). (1994). *A counselor's guide to career assessment instruments* (3rd ed.). Alexandria, VA: National Career Development Association.

Levinson, E. M., Ohlers, D. L., Caswell, S., & Kiewra, K. (1998). Six approaches to the assessment of career maturity. *Journal of Counseling & Development, 76,* 475–482. http://dx.doi.org/10.1002/j.1556-6676.1998.tb02707.x

Luyckx, K., Goossens, L., Soenens, B., & Beyers, W. (2006). Unpacking commitment and exploration: Preliminary validation of an integrative model of late adolescent identity formation. *Journal of Adolescence, 29,* 361–378. http://dx.doi.org/10.1016/j.adolescence.2005.03.008

Mitchell, L. K., & Krumboltz, J. D. (1996). Krumboltz's learning theory of career choice and counseling. In D. Brown & L. Brooks (Eds.), *Career choice and development* (3rd ed., pp. 233–280). San Francisco, CA: Jossey-Bass.

Murphy, K. A., Blustein, D. L., Bohlig, A. J., & Platt, M. G. (2010). The college-to-career transition: An exploration of emerging adulthood. *Journal of Counseling & Development, 88,* 174–181. http://dx.doi.org/10.1002/j.1556-6678.2010.tb00006.x

Nevill, D. D., & Super, D. E. (1986). *The Salience Inventory: Theory, application and research.* Palo Alto, CA: Consulting Psychologists Press.

Phillips, S. D., & Jome, L. M. (2005). Vocational choices: What do we know? What do we need to know? In W. B. Walsh & M. L. Savickas (Eds.), *Handbook of vocational psychology* (3rd ed., pp. 127–153). Mahwah, NJ: Erlbaum.

Porfeli, E. J., Lee, B., Vondracek, F. W., & Weigold, I. K. (2011). A multi-dimensional measure of vocational identity status. *Journal of Adolescence, 34,* 853–871. http://dx.doi.org/10.1016/j.adolescence.2011.02.001

Rounds, J. D. (1990). The comparative and combined utility of work value and interest date in career counseling with adults. *Journal of Vocational Behavior, 37,* 32–45. http://dx.doi.org/10.1016/0001-8791(90)90005-M

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55,* 68–78. http://dx.doi.org/10.1037/0003-066X.55.1.68

Santa Cruz County Regional Occupational Program. (n.d.). *Work Values Inventory.* Unpublished instrument. Retrieved from http://www.rop.santacruz.k12.ca.us/resources/career_planning/index.htm

Savickas, M. L. (2000, August). Building human strength: Career counseling's contribution to a taxonomy of positive psychology. In W. B. Walsh (Chair), *Fostering human strength: A counseling psychology perspective.* Symposium presented at the annual meeting of the American Psychological Association, Washington, DC.

Savickas, M. L., & Porfeli, E. J. (2012). Career Adapt-Abilities Scale: Construction, reliability, and measurement equivalence across 13 countries. *Journal of Vocational Behavior, 80,* 661–673. http://dx.doi.org/10.1016/j.jvb.2012.01.011

Seligman, L. (1994). *Developmental career counseling and assessment* (2nd ed.). Thousand Oaks, CA: Sage.

Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology. An introduction. *American Psychologist, 55,* 5–14. http://dx.doi.org/10.1037/0003-066X.55.1.5

Stumpf, S. A., Colarelli, S. M., & Hartman, K. (1983). Development of the Career Exploration Survey (CES). *Journal of Vocational Behavior, 22,* 191–226. http://dx.doi.org/10.1016/0001-8791(83)90028-3

Suen, H. K. (2015). *Super's Work Values Inventory—Revised (SWVI–r)* [Technical Brief]. Retrieved from https://www.kuder.com/research/technical-briefs/supers-work-values-inventory-r/

Super, D. E. (1963). Self-concepts in vocational development. In D. E. Super, R. Starshevsky, N. Matlin, & J. P. Jordaan (Eds.), *Career development: Self-concept theory* (pp. 17–32). New York, NY: College Entrance Examination Board.

Super, D. E. (1980). A life-span, life-space approach to career development. *Journal of Vocational Behavior, 16,* 282–298. http://dx.doi.org/10.1016/0001-8791(80)90056-1

Super, D. E. (1990). A life-span, life-space approach to career development. In D. Brown & L. Brooks (Eds.), *Career choice and development* (2nd ed., pp. 197–261). San Francisco, CA: Jossey-Bass.

Super, D. E., & Knasel, E. G. (1981). Career development in adulthood: Some theoretical problems and a possible solution. *British Journal of Guidance & Counselling, 9,* 194–201. http://dx.doi.org/10.1080/03069888108258214

Super, D. E., Savickas, M. L., & Super, C. M. (1996). The life-span, life-space approach to careers. In D. Brown & L. Brooks (Eds.), *Career choice and development* (3rd ed., pp. 121–178). San Francisco, CA: Jossey-Bass.

Sympson, S. (1999). *Validation of the Domain Specific Hope Scale: Exploring hope in life domains* (Unpublished doctoral dissertation). University of Kansas, Lawrence.

Van den Broeck, A., Ferris, D. L., Chang, C.-H., & Rosen, C. C. (2016). A review of self-determination theory's basic psychological needs at work. *Journal of Management, 42,* 1195–1229. http://dx.doi.org/10.1177/0149206316632058

Van den Broeck, A., Vansteenkiste, M., De Witte, H., Soenens, B., & Lens, W. (2010). Capturing autonomy, competence, and relatedness at work: Construction and initial validation of the Work-related Basic Need Satisfaction scale. *Journal of Occupational and Organizational Psychology, 83,* 981–1001. http://dx.doi.org/10.1348/096317909X481382

Weiss, D. J., Dawis, R. V., England, G. W., & Lofquist, L. H. (1967). Manual for the Minnesota Satisfaction Questionnaire. *Minnesota Studies in Vocational Rehabilitation, XXI.* Minneapolis: University of Minnesota.

Wrzesniewski, A., McCauley, C., Rozin, P., & Schwartz, B. (1997). Jobs, career, and callings: People's relations to their work. *Journal of Research in Personality, 31,* 21–33. http://dx.doi.org/10.1006/jrpe.1997.2162

Psychoanalytic Inquiry, 35:74–97, 2015

ISSN: 0735-1690 print/1940-9133 online
DOI: 10.1080/07351690.2015.987595

Routledge
Taylor & Francis Group

Quantifying Complexity: Personality Assessment and Its Relationship with Psychoanalysis

Anthony D. Bram, Ph.D. and Jed Yalof, Psy.D.

The fields of personality assessment and psychoanalysis have an entwined history and share much in common, notably an appreciation of the importance of understanding a person with complexity and depth, including the role of unconscious (or implicit) psychological processes. Personality assessment (or diagnostic psychological testing) offers a complement to psychoanalysis' primarily idiographic approach by integrating it with a nomothetic one; that is, applying quantitative methods to determine in what ways and to what extent a person is similar or different relative to normative data. It is surprising, then, that contemporary psychoanalysts are largely unfamiliar with the field of personality assessment and seldom refer their patients for evaluation to assist with diagnostic formulation and treatment planning. In this article, we offer practicing analysts (1) a general description of the ways that testing can assist diagnostically, (2) an introduction to categories of psychological tests that sample functioning under varying conditions or from different vantage points, (3) a survey of assessment research that has provided empirical validation of key psychoanalytic concepts, (4) a window into the assessment process as it is applied clinically, and (5) cases to illustrate when and to what benefit analysts might consider referrals for testing. Examples include use of testing in instances when a new patient reports a history of repeated treatment failures; when patient and analyst are embroiled in a protracted impasse; and when a fine-tuned assessment of analyzability is warranted.

The fields of personality assessment and psychoanalysis have much in common, notably a commitment to understanding people in-depth and conceptualizing treatment in individualized, meaningful ways. The two disciplines share the premise that there is more to understand about a person and their relationships than he or she is able to simply self-report: Sophisticated understanding encompasses taking into account and synthesizing both (1) conscious (or explicit or declarative) and (2) unconscious (or implicit or procedural) psychological processes. Such commonalities are not surprising because the early history of personality assessment (historically referred to as diagnostic psychological testing[1]), a subdiscipline of clinical psychology, was entwined with the psychoanalytic thinking that predominated at the time. This dates back to the

Anthony D. Bram, Ph.D., is in private practice and with the Cambridge Health Alliance/Harvard Medical School and Boston Psychoanalytic Society and Institute.

Jed Yalof, Psy.D., is affiliated with Immaculata University, Department of Graduate Psychology, Immaculata, Pennsylvania, and the Psychoanalytic Center of Philadelphia, as well as being in private practice.

[1]Although some make a distinction between *personality assessment* and *diagnostic psychological testing*, we use these two terms, along with *psychological testing* or *testing* or *assessment*, interchangeably in this article. Similarly, we refer to the psychologist conducting testing or assessment as the *examiner, diagnostician*, or *assessor*.

1940s, when the psychologist and analyst David Rapaport initiated a tradition at the Menninger Foundation, and subsequently at the Austen Riggs Center, of using a battery of psychological tests to operationalize and measure psychoanalytic constructs central to diagnostic understanding essential for treatment planning (Schafer, 2006). The typical battery was a thoughtful blending of cognitive (i.e., intellectual), neuropsychological (sorting test, memory test, nonverbal facial recognition), and projective personality testing that assessed ego functioning under varied conditions (Rapaport, Gill, and Schafer, 1968; see also Yalof, 2006). Psychoanalytically-informed practitioners of personality assessment continue to apply a similar battery today (e.g., Bram and Peebles, 2014; Yalof and Rosenstein, 2014).

It was then Rapaport's student, Roy Schafer, who continued to publish extensively on psychoanalytic ego psychological applications to diagnostic psychological testing (Schafer, 1954) and psychoanalysis (Schafer, 1967), and mentor psychologists who used testing as a method for understanding ego organization and personality functioning (see Blatt, 2006).[2] There has since remained a healthy cadre of psychoanalytic clinicians and writers who have maintained and refined the approaches of Rapaport and Schafer (e.g., Allison, Blatt, and Zimet, 1968; Bram and Peebles, 2014; Kleiger, 1999; Lerner, 1998; Tuber, 2012). Such contemporary psychoanalytic approaches to personality assessment continue to be well represented and valued, if not predominant, within the field's major professional organization, the Society for Personality Assessment, and its well-respected publication, *The Journal of Personality Assessment*.

Historically, personality assessment and psychoanalysis have had a synergistic, mutually informative relationship, and potential exists for this to be cultivated even further (e.g., Bram and Peebles, 2014; Lerner, 1998). Personality assessment also offers a complement to psychoanalysis' primarily idiographic approach by integrating it with a nomothetic one, that is, applying quantitative methods to determine in what ways and to what extent a person is similar or different relative to normative data.

Given the historical links and simpatico conceptual attitude shared by psychoanalytic theory and personality assessment, one might think that psychoanalysts and other analytic clinicians would often refer for diagnostic psychological testing, especially in the current climate where there is a premium on treatment efficiency and quantifiable outcomes. But, as we know from our clinical experience, this is definitely not the case. Yet, a carefully crafted assessment can inform decisions about, for example, whether and how to proceed when considering conversion to analysis, restarting a previously terminated treatment with a difficult patient, or understanding a protracted transference-countertransference impasse. As such, we pose this question: Can even a well-trained analyst learn anything valuable from this type of evaluation that supplements clinical formulations derived from interview and direct treatment experience? A humbling answer is provided by S. A. Appelbaum's (1977) research indicating that when experienced analysts had a test report available to them, but did not use it, they were less able to predict treatment outcomes than analysts who used the test data in conjunction with interview data.

Because many analysts today are not familiar with personality assessment/diagnostic psychological testing, our aims in this article are fivefold. First, we describe the clinical role of personality assessment and what it offers. We define what is meant by *diagnostic* in diagnostic psychological testing and clarify the types of questions that assessments are better and less

[2] Although Schafer's (2006) clinical and scholarly focus later shifted away from psychological testing toward broader theoretical and technical developments of psychoanalysis, he maintained a deep appreciation for the way in which his professional identity was shaped foundationally through his earlier diagnostic testing work under Rapaport's tutelage.

equipped to answer. Second, we acquaint readers with different types of psychological tests that sample functioning under varying conditions or from different vantage points. We summarize, briefly, the advantages and limitations of the different classes of tests, setting the stage for our point that it is the integration of data from tests with different properties that yields the most comprehensive and clinically-meaningful understanding of personality functioning. Third, we survey assessment research to offer analysts an appreciation of how the field has provided empirical validation for key psychoanalytic concepts. We believe that for psychoanalysis to remain viable in the current evidence-based treatment and training environments, it is crucial for analytic ideas to be supported as scientifically credible and for analysts to have some grasp of this evidence to help justify their work to critics and even to potential analytic patients when necessary (Huprich and Bornstein, 2015; this issue). Fourth, we offer a window into the assessment process as it is applied clinically. Fifth and finally, we present a series of cases to illustrate when and to what benefit analysts might consider referrals for diagnostic psychological testing.

THE CLINICAL ROLE OF PSYCHOLOGICAL TESTING: IN WHAT WAY IS IT "DIAGNOSTIC"?

There are multiple models of mental health diagnosis, each offering a different and potentially complementary perspective for understanding patients and their difficulties. Three such models are (1) *taxonomic symptom-focused*, (2) *case formulation*, and (3) *treatment-centered*, each with is own set of advantages and limitations (see Bram and Peebles, 2014). Briefly, the taxonomic symptom-focused model—entailing cataloguing overt symptoms into different classes of psychiatric disorders—is the well-known type featured in the *Diagnostic and Statistical Manual* (*DSM*; 5th ed., American Psychiatric Association, 2013) and *International Classification of Diseases and Related Health Problems* (*ICD*; 11th ed., World Health Organization, in press). Familiar to and valued by psychoanalysts, case formulation involves synthesis of an explanatory narrative that contextualizes presenting symptoms and problems in a person's developmental history, temperament, styles of coping/defense, and relational patterns, among other factors (e.g., McWilliams, 1999). The treatment-centered model places emphasis on mapping intrapsychic and interpersonal factors and processes, and conditions under which they vary, that underlie manifest symptoms and have important implications for psychotherapeutic intervention (Peebles, 2012). Examples of treatment-centered factors relate to alliance, transference, optimal treatment focus, and potential sources of resistance.

Psychological testing, especially as practiced by analysts or analytically-oriented assessors, is aimed at answering questions relevant to treatment-centered diagnosis. Although testing can occasionally contribute to arriving at a symptom-focused *DSM* or *ICD* diagnosis if certain symptom- and trait-based measures are included in the test battery,[3] the primary tool for such diagnostic purposes is a psychiatric interview focused on symptoms and history. Findings from testing can certainly be integrated into a case formulation narrative, but the primary method of data gathering for such formulation also involves detailed interview-based history-taking.

[3]The Rorschach can also make a contribution to when there are questions about complex differential diagnosis of *DSM* psychotic (Kleiger, 1999) or personality disorders (Huprich, 2006).

Although not widely known among practicing analysts, there are literatures on the applicability of personality assessment/psychological testing to various treatment-centered diagnostic questions. These include: analyzability (Peebles-Kleiger, Horwitz, Kleiger, and Waugaman, 2006); the presence, nature, and severity of thought disorder (Kleiger, 1999); clarification of factors underlying manifest symptoms, notably differentiating among contributions of structural weakness (ego deficit), characterological factors, intrapsychic conflict, and trauma (Bram and Peebles, 2014); anticipating alliance-related pitfalls and facilitating factors (Bram, 2010, 2013; Bram and Peebles, 2014); and predicting potentially problematic transference-countertransference paradigms (e.g., Bram, 2013; Schafer, 1954).

DIFFERENT CLASSES OF PSYCHOLOGICAL TESTS: COMPLEMENTARY VANTAGE POINTS, ASSESSING "CONDITIONS UNDER WHICH," AND A PARADIGM FOR INTEGRATION

A psychoanalytically-informed assessor recognizes the importance of understanding a person from different vantage points, as well as sampling functioning under varying environmental and psychological conditions. The different vantage points that can be integrated include those offered by measures that are: (1) self-report, (2) performance-based, (3) other-report (collateral-report), and (4) clinician-rated. Each of these approaches has advantages and limitations yet each provides a disparate, unique, and legitimate lens through which to view and understand a person (see Bornstein, 2007). The art and science of clinical personality assessment involves the conceptual integration of findings yielded by the different methodologies. Here, we briefly summarize the four broad classes of measures and then describe a clinical and research paradigm for making sense of convergences and incongruities among the different types of measures. These measures are summarized as follows (for more detail, see Bram and Peebles, 2014):

Self-report measures include questionnaires (e.g., Beck Depression Inventory-II; Beck, Steer, and Brown, 1996), personality inventories, (e.g., Personality Assessment Inventory [PAI]; Morey, 1991) and structured interviews (e.g., Structured Clinical Interview for *DSM*-IV [SCID]; First, Spitzer, Gibbon, and Williams, 1996). The patient rates or describes her own subjective experience of whatever is being assessed (e.g., symptoms, functioning, personality characteristics, perceptions of relationships).[4] Limitations of self-reports include their being subject to defensive distortion and that some psychological processes are not readily accessed consciously (e.g., McClelland et al., 1989). *Performance-based measures* sample a person's functioning and behavior *in vivo*. Rather than asking about a particular characteristic or capacity, there is a task that a person performs from which it is possible to infer something about that characteristic or capacity. Neuropsychological (e.g., Wisconsin Card Sorting Test; Heaton et al., 1993) and intelligence (e.g., Wechsler Adult Intelligence Scale—Fourth edition; Wechsler, 2008) tests are cognitive performance-based measures that are often integrated into personality assessment. Performance-based measures are also comprised of ambiguous-demand tasks that include

[4]Note that, compared to self-report questionnaires, many structured interviews such as the SCID also enable the clinician to draw conclusions based on other data (e.g., appearance, nonverbal cues, interpersonal style). Thus, compared to questionnaires, structured interviews are somewhat less susceptible to defensive distortion and impression management.

both *stimulus attribution* (e.g., Rorschach) and *constructive* (figure-drawing tasks) methods (Bornstein, 2007). Performance-based measures vary in the level of external structure, ambiguity, emotional stimulation, and relational/dynamic content enabling inferences about conditions under which functioning varies (see Bram and Peebles, 2014). In performance-based personality measures, the respondent's output can be rated according to criteria related to the construct being assessed. Compared to self-report, such measures tap more into unconscious or implicit psychological processes that are predictive of behaviors over the long term (McClelland et al., 1989), but are complicated and time-consuming to score, and therefore require careful review of cost-benefit ratio when using for clinical or research purposes. *Other-report measures* typically take the form of rating scales (e.g., Conners Revised Parent Rating Scale; Conners et al., 1998) completed by a spouse, parent, teacher, or other who knows the person well. Like self-report questionnaires, other-report measures are relatively easy, efficient, and inexpensive to administer, score, and interpret. Although they do not tap into one's subjectivity, they have the advantage of enabling assessment of how one's behavior is perceived by and impacts others. As will be elaborated in the following, the convergence or incongruity between self- and other-reports can be clinically meaningful. *Clinician-rated measures* are similar to other-report measures, but designed to systematically discipline and harness clinical judgment (e.g., Countertransference Questionnaire; Betan et al., 2005) or Q-sorting technique (e.g., Shedler-Westen Assessment Procedure; Shedler and Westen, 2007).

PARADIGM FOR INTEGRATING FINDINGS FROM DIFFERENT TYPES OF MEASURES

Making sense of the convergence or incongruity among findings from different types of measures is at the heart of a rich, psychoanalytically sophisticated assessment. This phenomenon has been studied empirically by psychoanalytically-minded researchers, mostly in the context of integrating self-report and performance-based findings. We present three lines of research that illustrate this nuanced, clinically-relevant paradigm, involving the assessment of (1) psychological health versus distress, (2) dependency, and (3) self-esteem/narcissistic vulnerability.

Assessment of Psychological Health Versus Distress

Mainstream mental health researchers and clinicians have, for purposes of diagnostic and outcome assessment, relied heavily on self-report questionnaires or interviews. Measures such as the Beck Depression Inventory-II (BDI-II; Beck et al., 1996) are prototypes of this common method. Appreciating the ubiquity of defensive processes, psychoanalysts have known, however, that there is more to the story than what the patient consciously and intentionally communicates. Psychoanalysts listen to the content of what their patients say (analogous to their self-report ratings) but, crucially, also attend to how they tell their story (Schafer, 1958), with attention to the content, emotional tone, word choice, syntax, etc. Shedler, Mayman, and Manis (1993) employed self-report measures in tandem with the narrative, performance-based Early Memories Test (EMT; Mayman, 1968) to demonstrate the clinical meaningfulness of the incongruities between the two types of measures. Notably, Shedler et al. identified a group of participants who self-reported low psychological distress yet showed signs of emotional disturbance in the content

and structure of their early memory narratives. These participants, whom the authors categorized as experiencing *illusory mental health*, showed greater physiological reactivity to various stress tests in the laboratory. This study was remarkable in demonstrating that there are real physiological costs to defenses. In a prospective study, the defensive style associated with illusory mental health has been shown to predict higher levels of health service utilization (Cousineau and Shedler, 2006).

Assessment of Dependency

Assessing a patient's dependent characteristics is essential in a psychoanalytic evaluation as this variable has implications for alliance, transference-countertransference paradigms, and establishing treatment focus. Bornstein's research (1999, 2002) has demonstrated that the thoughtful synthesis of self-report and performance-based data provides a more nuanced and accurate assessment of a person's dependency than use of either type of measure alone. The primary performance-based measure in this research has been the Rorschach Oral Dependency (ROD) scale (Masling, Rabie, and Blondheim, 1967), which scores the frequency or percentage of Rorschach responses that are marked by references to themes involving food, eating, passivity, helplessness, and other similarly regressive attitudes (see Bornstein and Masling, 2005)[5]. The self-report measure used in this research is often the Interpersonal Dependency Inventory (IDI; Hirschfeld et al., 1977). Thus, Bornstein classified respondents into four quadrants, based on whether they score (1) high versus low on the ROD and (2) high versus low on the IDI: Those high on both measures are said to have *high dependency*; those low on both, *low dependency*; those high on ROD but low on IDI, *unacknowledged dependency*; and those low on ROD but high on IDI, *dependent self-presentation* (Bornstein, 1998b). Help-seeking behavior, character style, and treatment implications will vary as a function of the quadrant in which a given person is located (see Bornstein, 1998a, 1998b).

Assessment of Self-Esteem/Narcissistic Vulnerability. Similar research methodology, employing both performance-based and self-report measures, has been applied to the study of self-esteem (e.g., Schroder-Abe, Rudolph, and Schutz, 2007). Explicit self-esteem has been measured with the self-report Multidimensional Self-Esteem Scale (MSES; Schutz and Sellin, 2006); and implicit self-esteem, by a version of the Implicit Association Test (IAT), a laboratory-based, computerized measure (Rudolph, Schroder, and Schutz, 2006, cited in Schroder-Abe et al., 2007).[6] Similar to Bornstein's (2002) previously mentioned paradigm, this design enabled research participants to be classified in four quadrants: *high self-esteem* (high scores on both the MSES and IAT), *low self-esteem* (low on both), *fragile self-esteem* (high on MSES, low on IAT), and *damaged self-esteem* (low on MSES, high on IAT). Schroeder-Abe et al. (2007) found that people with *fragile* or *damaged self-esteem* (compared to those with unequivocally high or low self-esteem) were more likely to suffer poorer mental and physical health.

[5]Recently, the ROD has been incorporated into the Rorschach Performance Assessment System (Meyer et al., 2011) where it has been renamed Oral Dependent Language.

[6]There are good reasons for the development of clinically pragmatic applications the IAT and/or for other performance-based measures of self-esteem. The self-esteem scale of the Social Cognition and Object Relations Scale-Global Rating Method (SCORS-G; Stein et al., 2011) based on the TAT or other narrative data holds promise for this. Unfortunately, the Rorschach Egocentricity Index, which may have been useful for this purpose, has recently been determined to be of questionable validity (Mihura et al., 2013).

ASSESSMENT RESEARCH AND KEY PSYCHOANALYTIC CONSTRUCTS

A longstanding, common criticism of psychoanalysis from academic and scientific circles has been that its constructs are unmeasurable and its tenets and hypotheses are untestable (e.g., see Yalof, 2015; this issue). In this section, we offer a survey[7] of measures of psychoanalytic concepts that have been researched and that have the potential to be selectively integrated into clinical personality assessment. Regarding the latter, we emphasize that it would be disingenuous to imply that all or most of these measures are, or should be, routinely applied in clinical assessment practice; especially the performance-based research measures that typically require extensive, specialized training to reliably score and interpret.[8]

Research Measures of Object Relations[9]

Object relations refer to internalized representations of self and other, that is, cognitive-affective relational templates or implicit relational expectations. Contemporary psychoanalytic theory holds that object relations are internalized based on early relationships with caregivers, and the resulting representations have crucial implications for ongoing interpersonal functioning, affect regulation, and thus overall adaptation and functioning (e.g., Bowlby, 1969; Fairbairn, 1952; Kernberg, 1976; Klein and Riviere, 1964; Winnicott, 1965).

Self-report questionnaires, notably the Bell Object Relations Inventory (BORI; Bell, Becker, and Billington, 1986; revised as the Bell Object Relations and Reality Testing Inventory, BORRTI; Bell, 1995) have shown good psychometric properties, relating in diagnostically meaningful ways to measures of interpersonal functioning and psychopathology (Hansen, 2001). The BORRTI yields scores on four object relations scales: Alienation, Insecure Attachment, Egocentricity, and Social Incompetence. Burns and Viglione (1996) developed a collateral-report version of the BORI, for ratings from the spouse's perspective (Spouse version of the BORI).

Because object relations—cognitive-affective relational templates—are embedded in procedural/implicit memory and, to a large extent, operate unconsciously, there are limits (e.g., defensive style, reflective capacity) to a person's ability to access and self-report them. Fortunately, there has been no shortage in the development of performance-based measures of object relations that are better able to tap into such implicit processes (see reviews by Lerner, 1998; Stricker and Healy, 1990). A number of such measures have been created employing traditional tools of psychological assessment, the Rorschach and Thematic Apperception Test (TAT; Murray, 1943). For the Rorschach, measures include Blatt, D'Affiti, and Quinlan's (1976) Concept of the Object on the Rorschach (COR) scale and Urist's (1977) Mutuality of Autonomy

---

[7] Space limitations preclude a comprehensive review.

[8] Rigorous training for scoring such measures is typically provided within research laboratories for specific projects. Such training is usually not available to, or practical for, most clinical assessors. It would be unusual if even a highly experienced clinical diagnostician had formal training in more than one of these research measures.

[9] Attachment theory overlaps with theories of object relations, though it diverges in important ways as well (see Fonagy, 2001). In the past two decades, attachment theory has been increasingly embraced by and stimulated creative research and clinical applications among analysts. Because of space limitations, we are unable to take up the measurement of attachment-related constructs. Suffice it to say that self-report (questionnaire and structured interview) and performance-based measures have been developed and studied extensively (see summaries by Fonagy, 2001; Stein et al., 1998; see also George and West, 2012).

(MOA) scale. Rooted in a Rapaportian approach to testing and ego psychology (Rapaport et al., 1968) as well as object relations (e.g,, Jacobson, 1964; Mahler, Pine, and Bergman, 1975), and cognitive developmental (Piaget, 1937; Werner, 1957) theories, the COR entails rating Rorschach response that involve quality, integration, and accuracy of human representations. Research with COR has demonstrated that Rorschachs of more severely disturbed patients show higher developmental scores on inaccurately perceived responses and lower developmental scores on accurately perceived responses (Blatt et al., 1976). Numerous other studies have demonstrated COR scores to (1) discriminate among diagnostic groups in theoretically consistent ways and (2) demonstrate and explain treatment efficacy in intensive inpatient and outpatient psychotherapeutic settings (summarized in Blatt, 2004).

Urist's (1977) MOA scale is another developmentally-based object relations coding scheme for the Rorschach. Unlike the COR, which restricts ratings to only human or human-like Rorschach responses (of which the capacity to construct is a developmental accomplishment in and of itself, which some patients have not achieved) the MOA can be scored to assess "thematic content of relationships (stated or implied) between animal, inanimate, or human perceptions on the Rorschach" (Monroe et al., 2013, p. 538). A recent meta-analysis of research on the MOA supported its impressive validity including its success at "capturing group differences . . . between clinical and nonclinical populations, diagnostic group differences, and differences between groupings based on behavioral criteria (e.g., self-mutilators vs. non-self-mutilators). . . . [And at] discerning discrete behavioral markers (e.g., number of psychotherapy sessions attended), psychotherapy outcome change data and level of symptomatology/overall functioning" (Monroe et al., 2013, p. 552). Although, to date, MOA has not yet been widely applied clinically (only 13% in a recent survey of experienced assessors; Meyer et al., 2013), a version has recently been included as part of the Rorschach Performance Assessment System (Meyer et al., 2011) designed for clinical use, so this measure of object relations is expected to become more integrated into routine, real-world testing practice.

The Social Cognition and Object Relations Scale (SCORS; Stein et al., 2011; Westen, 1995; Westen et al., 1985) is a performance-based measure of object relations that can be applied to the TAT and other narrative data (e.g., EMT, therapy transcripts). Westen and colleagues developed the SCORS based on integration of principles from object relations theories with ideas and methods from social cognition theory based in academic psychology (e.g., Worchel, Cooper, and Goethals, 1991).

The most recent version, the SCORS-Global Rating Method (SCORS-G; Stein et al., 2011) enables TAT stories (or other narratives) to be rated on scales for eight dimensions of object relations: (a) complexity of representations of people, (b) affective quality of representations, (c) emotional investment in relationships, (d) emotional investment in values and moral standards, (e) understanding of social causality, (f) experience and management of aggressive impulses, (g) self-esteem, and (h) identity and coherence of self. Studies of various iterations of the SCORS have shown support for the association between internal object relations and disturbances in interpersonal and adaptive functioning (e.g., Ackerman et al., 1999; Bram, 2014). Support has also been found for psychoanalytically-informed hypotheses of developmental antecedents, especially involving trauma, of disturbed object relations (e.g., Bram, Gallant, and Segrin, 1999; Kernhof, Kaufhold, and Grabhorn, 2008; Slavin et al., 2007). Kelly (1997) has described treatment planning implications of the SCORS (as well as the Rorschach MOA) in work with adolescents,

though there is much room for more writing on its clinical application. With publication of normative data for the SCORS-G and further dissemination of training (Bram, 2014), this measure has potential for wider clinical application, as Westen (1995) originally intended.

A final performance-based measure of object relations to be discussed is one that does not require administration by an assessment psychologist and, thus, potentially could be applied in practice by nonpsychologist clinicians:[10] The Object Relations Inventory (ORI; see Blatt, 2004) requires respondents to write or verbalize descriptions of self and significant others (e.g., mother, father, spouse, therapist), with key adjectives queried for further elaboration. The ORI can be scored on various thematic dimensions (traits related to benevolence, punitiveness, ambition, and ambivalence) as well as structural scales assessing conceptual level (i.e., degree level of cognitive development within object representations, ranging from a sensorimotor-preoperational level in which people are described solely in terms of their selfobject functions to the most mature level in which descriptions are complex but well-integrated) and differentiation-relatedness (ranging from representations that reflect compromise in self-other boundaries to those that depict creatively integrated, reciprocal relationships). There is extensive empirical support for these ORI scales (summarized in Blatt, 2004).

Research Measures of Defenses

For most psychoanalysts, a thorough diagnostic understanding accounts for a patient's defensive organization. A number of self-report scales, which are easily administered and scored, have been developed. For example, the Defense Style Questionnaire (DSQ; Andrews, Singh, and Bond, 1993) is a 40-item[11] self-report scale, popular in research, in which items tap into, and enable scoring on, scales assessing defenses that are considered mature (e.g., anticipation, humor, sublimation), neurotic (e.g., idealization, reaction formation, undoing), or immature (e.g., acting out, denial, projection, splitting). The DSQ has good psychometric properties and, importantly, its authors of have published norms that can facilitate interpretation when applied clinically (Andrews et al., 1993). Performance-based measures of defenses have been developed involving rating Rorschach and TAT responses. Cooper and Arnow's (1986) Rorschach Defense Scales (RDS) systematize, clarify, and illustrate many of Schafer's (1954) seminal ideas about assessment of neurotic- and psychotic-level defenses but also, consistent with Kernberg's (1967) later ideas about level of personality organization, conceptualize defenses at the borderline level, as well. The RDS enables ratings on 15 different defenses. To offer a sense of the scoring, the borderline-level defense of splitting is recorded for Rorschach responses such as "A person being torn between two sides, one good and one evil, and they're both pulling against each other" and "That looks like a river with the upper half being majestic and divine and the other half as being evil and sinister" (Cooper and Arnow, 1986, p. 21). The specificity of the RDS in being able to identify defenses makes it of potential appeal to assessment practitioners but also makes reliable scoring challenging, at least for certain defenses (see Table 1 in Cooper, Perry, and O'Connell, 1991).

[10]This would require arrangements with a trained rater to score. There is presently a paucity of such raters available.

[11]Note that there are multiple versions of the DSQ with varying numbers of items.

For the TAT and similar storytelling tasks, Cramer (1991) has developed the Defense Mechanisms Manual (DMM) that enables rating of narratives on three types of defenses, from least to most developmentally mature: denial, projection, and identification. The DMM has been extensively researched, empirically supporting numerous psychoanalytic hypotheses, including that with development children exhibit more mature defenses, that stress is associated with increased defensive operations, that defenses are associated in predicted ways with personality functioning and psychopathology, and that defenses can be modified psychotherapeutically (summarized in Cramer, 2006).

Research Measures of Mentalization

Over the past two decades, perhaps the most generative new concept in psychoanalysis—from both clinical and research standpoints—has been that of mentalization. Emanating from an integration of frameworks from attachment theory, psychoanalysis, philosophy, and developmental cognitive neuroscience, mentalization refers to the capacity to think about one's own and others' internal motivations and psychological states (thoughts, feelings, desires, etc.; Fonagy et al., 2002). A child's capacity to mentalize is conceptualized as developing out of a secure attachment (marked by affective attunement and accurate mirroring) with caregivers and having critical implications for the maturation of adaptive relational capacities and affect regulation (Fonagy et al., 2002) To date, mentalization has been predominantly assessed by psychoanalytic clinical researchers using the Reflective Function (RF) scale, a performance-based measure, as applied to Adult Attachment Interview (AAI) narratives (Fonagy et al., 1998). Similar to the AAI itself,[12] the training program to reliably score the RF is extensive and takes considerable time and practice, so thus far has not been conducive to routine clinical application (Conklin, Malone, and Fowler, 2012).

There also is a promising new performance-based measure of *mental state discourse*, closely related to mentalization, that can be applied to storytelling tasks such as the TAT (Symons et al., 2005). Symons et al.'s measure is relatively straightforward to learn to score reliably: It involves tallying references within narratives to cognitive states (e.g., think, believe, remember), affective desire states (e.g., want, desire, wish), and affective states (e.g., afraid, hate, love). This measure has garnered little clinical research thus far, save for a study indicating that mental state discourse differentiated between preadolescent girls at high versus low risk for developing an eating disorder (Cate et al., 2011). Although further research on its psychometric properties is needed, and normative data would need to be collected and disseminated to optimize its interpretability clinically, this measure of mental state discourse has potential for application in clinical practice: It can be applied to an instrument already in common use (TAT) and scoring is relatively uncomplicated and efficient, not requiring extensive additional training (Cate et al., 2011). Similarly promising for research and clinical assessment of mentalization is an approach to conceptualizing constellations of extant Rorschach Comprehensive System (Exner, 2003) variables recently proposed by Conklin, Malone, and Fowler (2012). Specifically, Conklin et al. presented a model for assessing mentalization based on the extent of a patient's use of shading as texture (conceptually

[12] An abbreviated version of the Adult Attachment Interview (AAI) has been recently developed to promote greater efficiency in assessing RF (Falkenstrom et al., 2014).

and empirically related to attachment need), human movement responses (related to empathic capacity), and ratio of *good human* versus *poor human* responses (related to quality of object representations).

Research on Other Psychoanalytically Relevant Diagnostic Considerations

Psychoanalysis has offered clinically meaningful frameworks and useful methodologies to understand patients diagnostically that can be used to complement and enrich mainstream symptom-based psychiatric approaches (American Psychiatric Association, 2013). Here, we briefly describe approaches to assess psychoanalytic concepts that have crucial treatment implications including (1) level of personality organization (e.g., Kernberg, 1967), (2) the distinction between anaclitic and introjective personality and pathology (Blatt, 2004), and (3) personality disorders (Shedler and Westen, 2007).

Assessment of level of personality organization. Kernberg's (1967) structural approach to diagnostic conceptualization in terms of level of personality organization (neurotic vs. borderline vs. psychotic) has proved useful to clinicians in planning treatment (McWilliams, 2011) and has lent itself well to empirical assessment.[13] Kernberg and colleagues have developed both a semistructured interview (Clarkin et al., 2004) and a self-report measure (Lensenweger et al., 2001) designed to assess ego capacities (e.g., identity cohesion, object relations, defensive organization, reality testing) crucial to diagnosing level of personality organization. The Structured Interview of Personality Organization (STIPO; Clarkin et al., 2004), designed to be administered by experienced clinicians, involves detailed inquiry into a patient's behavior and inner experience, which are subsequently rated by clinicians in terms of each of the ego capacities. The STIPO blends self-report and performance-based/clinician-rating methodology, as the interviewer or rater makes scoring judgments based on the content and process of the interview.

The Inventory of Personality Organization (IPO; Lensenweger et al., 2001; 57 items) and its shortened revision (Inventory of Personality Organization-Revised; IPO-R; Smits et al., 2009; 41 items) are self-report measures that, compared to the STIPO, more efficiently, albeit with less descriptive richness and without expert clinical judgment, tap into similar ego capacities associated with differential diagnosis among levels of personality organization.

Assessment of introjective versus anaclitic types of personality and pathology. Blatt's (2004) conceptualization that personality and vulnerability to certain types of psychopathology can be understood in terms of anaclitic (predominant concerns about relatedness involving themes of dependency, separation, and loss) versus introjective (concerns around self-definition involving themes of self-criticism, perfectionism, achievement) styles (Blatt, 2004). Blatt's research has shown that patients with each of these types of personality preoccupations respond variably to different types of psychotherapy in theoretically expected ways in terms of alliance formation and outcome (e.g., patients who are anaclitic respond positively to brief, more supportive treatments; those who are introjective require longer-term, insight-oriented/expressive treatments with a greater transference focus; summarized in Blatt, 2004).

[13]Level of personality organization is represented by the P Axis in the Psychodynamic Diagnostic Manual (PDM; PDM Task Force, 2006).

Blatt and colleagues (Blatt, D'Affiti, and Quinlan, 1976) developed the Depressive Experiences Questionnaire (DEQ), a 66-item self-report measure enabling assessment on anaclitic and introjective dimensions. The DEQ has been used extensively in research, lending empirical support to numerous psychoanalytic hypotheses (summarized in Blatt, 2004). One criticism of the DEQ has been its complex scoring algorithm that, among other things, limits its efficient clinical application (e.g., Flett et al., 1995). Thus, various revised versions of the DEQ have been developed that are both shorter and scored more straightforwardly (summarized in Desmat et al., 2007). Desmat et al. (2007) compared the original version and its revisions and determined that a more easily scored 19-item version (Bagby et al., 1994) possessed optimal psychometric properties; such a briefer and more simply scored version is more realistic to integrate into clinical assessment.

Assessment of personality disorders. There are various ways to assess personality disorders, including the utilization of the Rorschach test with other measures (e.g., Huprich, 2006). A psychoanalytically sophisticated, clinician-rated measure of personality and personality pathology is the Shedler-Westen Assessment Procedure (SWAP; e.g., Shedler and Westen, 2007). The SWAP involves a clinician's (computerized or manual) sorting of 200 statements, each describing personality and interpersonal processes, into eight categories from least to most descriptive (the more descriptive, the more heavily rated). A clinician can complete a SWAP based on at least six therapy sessions with a patient or a single intensive, semistructured interview (Westen, 2004; Westen and Muderrisoglu, 2003). Items on the SWAP include experience-near descriptions of defenses and object relations (e.g., "Is quick to assume that others wish to harm or take advantage of him/her; tends to perceive malevolent intentions in others' words and actions"). The SWAP yields scores on scales associated with various personality disorders, as well as a narrative description of the patient. For clinical assessors, a SWAP can be included in a test battery, either by having the sort completed by the referring clinician (assuming that clinician had the necessary contact with the patient) or by the assessor him/herself (based on conducting the required interview).

THE CLINICAL TEST BATTERY

A battery of tests, rather than reliance on a single measure, facilitates sampling—and thus inferences about—the patient's functioning under varied environmental and psychological conditions, an understanding of which is essential in answering referral questions (Bram and Peebles, 2014; Rapaport et al., 1968). Through selection of tests in the battery, conditions can be varied in terms of (1) level of external structure of tasks (how clear and explicit expectations and guidelines are versus how much on his/her own a patient is to figure out an appropriate response), (2) degree and type of emotional stimulation, and (3) relational-dynamic themes evoked (e.g., dependency, aggression, sexuality, etc.; Bram and Peebles, 2014). Thus, in a relatively brief period of time, the diagnostician can sample the patient's functioning under meaningfully varied conditions by integrating formal scores, response content, the patient' test-taking attitude and behavior, and the data from the patient-examiner interaction (including the examiner's countertransference). Regarding the latter, the psychoanalytic diagnostician conceptualizes the relationship with the patient as a *screen test* for psychotherapy, that is, can systematically test various hypotheses about what

facilitates and impedes engagement and collaboration and make inferences about what types of transference-countertransference configurations are apt to unfold psychotherapeutically (Bram, 2013; Bram and Peebles, 2014; Schafer, 1954; Shectman and Harty, 1986). The methodology described earlier, involving examining convergences and incongruities among different sources of data, contributes to multilayered, *conditions under which* inferences that result in sophisticated, treatment-relevant formulations regarding psychic structure/ego functioning (e.g., reality testing, logical reasoning, affect regulation), object relations (implicit relational expectations), and maturity of self-development.

The diagnostician tailors the test battery depending on what will best answer the referral questions (Bram and Peebles, 2014; Meyer et al., 2001). That said, psychoanalytic diagnosticians often build off of (or subtract from) a core of the traditional Menninger battery (Rapaport et al.,1968) entailing the Rorschach, TAT, the Wechsler intelligence test, adding a self-report personality inventory such as the Minnesota Multiphasic Personality Inventory-2nd edition (MMPI-2; Butcher et al., 1989) or PAI and, when relevant, collateral-report measures. There are times when some of the previously mentioned research measures are also included in the clinical battery. When there are specific questions about autonomous ego functions related to memory, executive functioning, learning disability, neuropsychological and psychoeducational tests can be integrated.[14]

PSYCHOANALYSTS AND THE REFERRAL FOR PSYCHOLOGICAL TESTING: CASE ILLUSTRATIONS

There are a multitude of scenarios in which a referral for diagnostic psychological testing can be valuable to psychoanalysts and their patients. Here, we illustrate three such common clinical scenarios: (1) planning treatment with a new patient who reports a history of treatment failures, (2) understanding a current therapeutic impasse, and (3) when there are questions about analyzability.

Referral for Testing: New Patients with Past Treatment Failures

One occasion an analyst might consider a referral for testing is when a new patient comes for consultation describing a history of treatment failures. Consider the case of Betsy, a young adolescent whose symptoms included chronic depression, self-harm, oppositionality, declining academic functioning, and multiple somatic complaints (without medical explanation) resulting in frequent absences from school (see Bram, 2010; Bram and Peebles, 2014). Betsy's parents brought her to a psychiatrist-analyst for pharmacotherapy consultation. The psychiatrist was struck not only by the patient's refusal to engage in give-and-take in the consultation but by her parents' description of a series of previous unsuccessful psychotherapeutic and pharmacological efforts, often involving Betsy's literally walking out. The parents also related that clinicians from different mental health disciplines had offered a confusing, conflicting array of descriptive *DSM* diagnoses: oppositional-defiant, Asperger's, major depression, generalized anxiety, and bipolar. Despite the

[14]When the diagnostician does not have specialized expertise in these areas, he or she partners with a colleague who does.

family's desperation for immediate intervention with a new prescription, the consultant humbly and wisely declined. Instead, she referred for assessment aimed at answering the two primary questions: (1) what has impeded Betsy's ability to form a therapeutic alliance (the strongest predictor of therapeutic outcome; Horvath, Del Re, Fluckiger, and Symonds, 2011) and what conditions might enhance or impede such capacity? and (2) what is the underlying developmental disruption (the role of structural weakness/ego deficit, character, conflict, and/or trauma; see Chapters 8 and 9 in Bram and Peebles, 2014, and Chapter 11 in Peebles, 2012) driving her symptoms and that can guide treatment focus? Regarding the latter, the psychiatrist wanted to rule out an undiagnosed thought disorder, a specific type of structural weakness.

Betsy reluctantly engaged in testing with an analyst-diagnostician, but it was the process of working with her resistance—part of the patient-examiner data—that contributed to formulating answers about what would facilitate versus impede an alliance should she be referred for another attempt at psychotherapy. The test data underscored that her underlying developmental disruption involved a combination of structural weaknesses—in the ego functions of affect regulation and logical reasoning—and characterological disturbance involving elements of both hypervigilant and oblivious narcissism (Gabbard, 1989).

To offer a sense of how test data can be thoughtfully assembled to provide a meaningful and elaborated understanding, consider the following data points contributing to inferences about her affect regulation. Structural scores on the Rorschach that could be compared to nonpatient norms highlighted both (1) her efforts to emotionally constrict (low overall response output and low responsiveness to the more affect-laden fully chromatic cards) and (2) the failure of such defensive efforts as manifested by elevated scores on an index showing the breakthrough of primary process themes. This converged with TAT stories involving implicit premises of emotional dysregulation and challenges containing feelings of sadness and loss, which are minimized and morph into and erupt in anger and aggression. These data, along with other convergent evidence illuminated how difficult—painful, helpless, overwhelming—it must have been for Betsy to sit with a therapist with the expectation that she would access, tolerate, and verbalize her internal experience.

Similarly, her test responses and scores offered a glimpse into her internal object world and thus deepened an understanding of what made a therapeutic alliance with Betsy so challenging and elusive. Rorschach responses such as an animal "killed for sport," "a griffin about to kill," and a "bear . . . kicked out because it doesn't fit in" capture evocatively her malevolent, hostile, rejecting, and unforgiving internal objects and relational expectations. Moreover, her provocative behaviors—manifested during the evaluation in efforts to challenge the examiner's authority and making snide and sarcastic comments—rendered her likely to actualize and reinforce her dismal relational expectations through enactment or projective identification (Gabbard, 1995). This returns us to the primary referral question: What would it take to connect with her? Through a series of disciplined hypothesis-testing interventions within the patient-examiner relationship (Bram and Peebles, 2014) aimed at assessing what would facilitate collaboration to complete the evaluation and begin to open up, the following was learned: Collaboration could be promoted by (1) anticipating interpersonal provocations but, whenever possible, containing projective identifications rather than responding in kind; (2) not pushing for reflectiveness, especially at first; (3) engaging around conflict-free topics (e.g., musical interests, movies), which may enable communication of concerns through displacement; (4) refraining from interpretation; and (5) giving her the space to self-regulate distance/closeness. This provided a map for a particular type of

psychoanalytic psychotherapy on the supportive end (A. H. Appelbaum, 1989) of the supportive-expressive continuum (Wallerstein, 1986). Although initially rejecting a recommendation for therapy, after further postevaluation struggles, she requested seeing the diagnostician-analyst for psychotherapy. Applying such findings and implications from the psychological testing created the conditions under which Betsy could engage in a three-year psychotherapy in which her emotional and interpersonal functioning dramatically improved (see Bram, 2010, for details of the treatment and outcome).

Referral for Testing: Treatment Impasses

When embroiled in a treatment impasse, analysts seek consultation with a trusted mentor or colleague. One type of consultation can be obtained through a referral to an analytic colleague who conducts psychological testing. This type of consultation can be particularly useful when there is a question about whether the impasse may be the result of missing something diagnostically (e.g., Might the patient suffer a structural weakness in reasoning and reality testing that has not been previously identified? a characterological disturbance not evident earlier in treatment?).

Consider the case of Dr. A, an advanced psychoanalytic candidate, who was struggling in the second year of his twice-weekly treatment of Dave, a depressed, inhibited middle-aged man. Dr A found himself preoccupied with wanting to terminate treatment and refer Dave on "because someone else, probably a senior colleague, would do a better job." In response to a sense that Dave was stagnating and hopeless, Dr. A felt therapeutically impotent. Consulting with a mentor whom Dr. A fantasized would agree to take the patient herself, the mentor instead suggested that a referral for testing might illuminate the dynamics creating the impasse. Dr. A made the referral for testing, but with a sense of dreadful shame and deficiency because "I can't crack this patient's code."

In collaboration with the assessor, Dr. A formulated referral questions related to clarifying the patient's underlying developmental disruption—in particular, could there be something going on characterologically that was making it so hard for Dave to take in and accept what Dr. A had to offer and, in turn, for Dr. A to feel efficacious? Test scores did not point to Dave's having any major structural weakness: On the Rorschach, for example, structural scores that could be compared to normative data, did not reveal pervasive ego deficits in affect regulation, reality testing, or reasoning. What did become clear was that Dave's character style was marked by not only masochistic/self-defeating elements but sadistic ones as well. More specifically, testing clarified that this style was his best effort to manage unacceptable and frightening angry and aggressive impulses.

How did the diagnostician arrive at this formulation? Some clues were gleaned from the patient–examiner relationship. Persistently throughout the evaluation, Dave adopted an ostensibly kidding attitude toward the diagnostician (e.g., "I can't imagine how someone would want to write all of this down," "You really like doing this work?") that left the diagnostician feeling devalued himself and, at times, even a bit deskilled. Dave's devaluing attitude was interwoven with more self-critical and self-defeating comments. The latter converged with his profile on the self-report MMPI-2, but that profile did not reveal anything that fit that would account for Dave's subtly devaluing attitude. Dave's initial responses to the TAT and Rorschach were highly constricted. He offered barebones responses that were minimally revealing. He did not provide a sufficient number of responses for a valid Rorschach protocol, so he was instructed that the

cards needed to be readministered (Exner, 2003) because "we need more responses." Although he responded apologetically and self-critically ("I guess I blew it") revealing his harsh superego (see Yalof and Rosenstein, 2014), in response to the readministration instructions, his new thematic content illuminated the angry and aggressive impulses that he typically kept out of conscious awareness and were only expressed covertly: Rorschach content included, for example, a "fighter jet," "battleship firing torpedoes," and "mushroom cloud." Such content was often in configuration with structural scores associated empirically with an internal sense of feeling helpless and out of control. This aided the diagnostician to empathize with why it must be so difficult for Dave to consciously acknowledge and contain his anger and aggression. Moreover, the emergence of aggressive themes in the readministration were understood as reflecting both (1) his rage in response to the narcissistic injury (Kohut, 1972) of being told his initial responses were insufficient and (2) the emergence of typically defended-against expression of anger.

The resultant diagnostic reformulation was that Dave's entrenched depression and therapeutic impasse were manifestations of his underlying masochistic character style that was in place to manage intrapsychic conflict about consciously unacceptable and frightening aggressive impulses. This reformulation entailed crucial treatment implications. First, these findings enabled the diagnostician to reassure Dr. A that these dynamics would inevitably play out with any therapist: What was unfolding in the transference and countertransference was the heart of what needed to be addressed—rather than sidestepped—in treatment. Second, along similar lines, alerting Dr. A that the focus of treatment would need to be Dave's underlying developmental disruption (Bram and Peebles, 2014; Peebles, 2012) involving the central role of character underscored that treatment would require patience and time (Schlesinger, 1995). Third, the diagnostician explained to Dr. A that his experience of hopelessness is not an uncommon reaction in clinicians when working with patients who present masochistically (S. A. Appelbaum, 1963) but can be hard to recognize because of the rigidity of impasse and the associated countertransference feelings of frustration, guilt, and anger emerging from a projective identification. Fourth, the findings assisted Dr. A to recognize the importance of addressing the sadism that subtly pervades the transference and countertransference with patient's organized masochistically (McWilliams, 2011). Related to the latter, it was suggested respectfully that, given the understandable challenges that Dave's treatment posed, that Dr. A consider seeking supervision, with careful attention to process material and transference-countertransference dynamics, with a trusted analyst with whom it would feel safe to share and reflect on angry, hateful, and sadistic countertransferences (e.g., including wishes to be rid of the patient). Finally, the test findings helped rekindle Dr. A's empathy for Dave, especially with regard to better appreciating the function his masochistic style served, (i.e., to defend against frightening, intolerable aggressive feelings and impulses). This resonated clinically for Dr. A, as he understood that Dave grew up with a mother who was prone to rages herself, and had little tolerance for her children's expression of aversive feelings themselves.

Referral for Testing: Questions About Analyzability

Historically, the question of analyzability has been thorny one, typically posed as "To analyze or not?" with stringent criteria requiring neurotic-level personality organization and significant ego strengths (e.g., Bachrach and Leaff, 1978; Freud, 1904). As the field of psychoanalysis has matured, not to mention encountered a changing mental health environment, however, analysts

have increasingly appreciated the *widening scope* of patients who might benefit from intensive analytic treatment (Stone, 1954) and have adapted technique accordingly (e.g., Eissler, 1953; Karon, 2002; Kohut, 1984). Peebles-Kleiger et al. (2006) articulated that this shift in perspective enables the following reformulation regarding assessing analyzability: "In contrast to the traditional approach of rendering a thumbs-up or thumbs-down decision of accept or reject for analysis, the model we espouse considers *conditions under which* a meaningful and productive engagement can occur" (p. 505).

An example of the use of testing to answer questions related to analyzability involves a child psychiatrist-analyst, Dr. B, who had recently begun treating 16-year-old Elena in combined psychotherapy-pharmacotherapy for a myriad of refractory symptoms including inattention, lying, stealing, defiance, and academic and peer difficulties. Born in an Asian country, Elena was abandoned as an infant and spent three years in an orphanage before being adopted by American parents. Elena received intensive multidisciplinary services from the time of her adoption to address physical and neurodevelopmental delays. Although such services enabled her to catch up developmentally in important ways, the aforementioned symptoms lingered and then exacerbated as she entered high school. Until her referral to Dr. B, Elena's psychological and pharmacological treatment had been exclusively behavior and symptom focused. In their weekly sessions, Dr. B was struck by Elena's politeness and compliance that were at odds with reports of her significant behavior problems. Dr. B wondered if it might require a more intensive treatment to reach Elena and meaningfully impact what were likely underlying structural weaknesses and internalized object relations that gave rise to her symptoms. Dr. B also noticed that Elena's verbalizations were often confusing to follow and thus wondered whether Elena might suffer a thought disorder and might not be able to tolerate analysis without becoming cognitively disorganized and regressed. Dr. B referred Elena for testing to clarify (1) to what extent and how her severe behavioral problems might be underpinned by a disturbance in her reasoning and (2) whether, and under what conditions, she might benefit from analysis.

Although Elena's Rorschach and other performance-based measures indicated that she did not exhibit disordered thinking commensurate with an incipient psychotic illness (i.e., along the lines of schizophrenic- or bipolar-spectrum illness), the data suggested that she was vulnerable to moments of (1) highly confused thinking and illogical reasoning and (2) viewing people and the world in distorted ways. The diagnostician's careful attention to the configuration of thematic content and formal scores (Bram and Peebles, 2014; Peebles-Kleiger, 2002; Schafer, 1954) indicated that lapses in these ego functions were most associated with conditions of heightened emotion; arousal of her core sense of badness or longings for caretaking; and being more on her own with less external structure.

Rorschach and TAT data indicated that Elena suffered a profoundly damaged sense of self: Implicitly she experienced herself as bad, inadequate, disconnected, unstable, and lost. Her fundamental template for close relationships was that her needs are too onerous for others. Consider the following evocative and highly unusual response to a TAT card depicting a "woman's head against a man's shoulder" (Murray, 1943, p. 19):

> Johnny fell asleep in the car on his way home from school. And he came home from school really late. (Characters thinking and feeling?) His dad was about to carry him from the car. Um, Johnny is feeling really tired. And the dad is thinking "I really don't want to carry him." So he drops him right there on the floor. (And?) Johnny gets up and puts himself to bed.

Notice how Elena's distorted reality testing—seeing a parent–child relationship instead of two adults—is colored by the powerful relational expectation that a child's needs will burden the caregiver and result in the child's being "dropped."

Additionally, data from the performance-based Rorschach and self-report Minnesota Multiphasic Personality Inventory-Adolescent version (MMPI-A; Butcher et al., 1992) converged to indicate that Elena's efforts to manage her insecurities about herself and relationships involve a certain narcissistic solution wherein she minimizes vulnerabilities, inflates herself to peers, and portrays herself as self-sufficient. Also of note, scoring Elena's Rorschach, it became apparent that she was impacted by more emotions—especially aggression—than she was consciously aware of and able to articulate.

The diagnostician concluded that Elena was in need of an intensive psychotherapeutic intervention over the long-term. In his report, he articulated the following treatment implications:

> Brief and primarily symptom-focused therapies are unlikely to help her shift her entrenched, implicit, maladaptive views of herself and others. Analysis offers the best hope to help her reflect on and gradually revise her internal templates for relationships. She would benefit from a therapist who can be a stable, consistent, real presence in her life over time. . . . The test data indicate that her reactions and other behavior are often impacted by emotions of which she is not aware, so helping her to have a better handle on these internal experiences is crucial to help her guide decisions and regulate her behavior. . . . With its frequency over the long-term, analysis has the unique advantage of allowing her to present herself as she really is and play out her relational patterns in real time with the analyst who can help her observe and understand what she is doing. . . . With time, her mistrust and expectation of being "dropped" (abandoned) is likely to color her view of the therapist, and she is apt to become more anxious, manifested in such ways as becoming more withdrawn/dismissive/self-sufficient or acting to provoke her own rejection. This may be more likely in the context of vacations and other separations.

The diagnostician clarified further that if analysis is undertaken,

> it will be crucial for the analyst to recognize that this is a "widening scope" case in which Elena's structural weaknesses in reasoning, reality testing, and emotional regulation need to be taken into account. Optimally, verbal interventions will be brief, simple, and close to her experience (i.e., not deep or complex interpretations). At times, it may make sense to ask her to repeat back what she has understood the analyst to say to make sure that she has accurately taken in and processed accurately. Additionally, it will be important that the therapist attend to the pace of her emotional processing, respect her defenses, and sometimes help regulate the pace if she appears to be getting into themes or emotions that may be difficult for her to contain (e.g., if too close to the end of sessions or prior to an interruption). If Elena's communication becomes more confused, the analyst can use this as a clue that she may be emotionally overwhelmed, possibly related to a self-experience of badness and/or anxiety/insecurity about her relationship with the analyst.

The feedback from testing set the stage for what would be a successful three-year analysis at a frequency of 3–4 times weekly. Years later, Dr. B recalled the assessment as a "valuable second opinion . . . from a different frame of reference" that supported her conviction that analysis was the treatment of choice. Dr. B added that the reliability and validity of the tests in the battery gave her confidence in the findings and conviction in the recommendation for analysis that she would not have had otherwise. Dr. B noted that throughout the analysis, she held in mind the recommendation related to Elena's vulnerability to confusion and thus carefully attended to the brevity and clarity of her own therapeutic communications.

DISCUSSION

The role and function served by a psychoanalytically informed approach to diagnostic psychological testing has been understated in the professional, including psychoanalytic, literature (Bram and Peebles, 2014). Such understatement is surprising, given that the American Psychoanalytic Association's (2006) practice guidelines for clinical assessment includes the following statement:

> The utilization of psychoanalytically-oriented psychological testing has been shown to enhance and sharpen the psychoanalytic assessment process in three areas: (1) the assessment of analyzability, (2) the prediction of treatment outcome, and (3) the delineation of dimensions of change (or variables) by which treatment outcome may be measured (S. A. Appelbaum, 1976; Wallerstein, 1986). Due to the scarcity of this resource, it has been part of the psychoanalytic assessment process in only a few practice settings. Continued positive results from the use of this testing in these settings might lead to greater availability of this resource and support for its wider use. [p. 4]

We believe that even experienced psychoanalysts can benefit from collaboration with psychological assessors and, in particular, that such psychodiagnostic consultations can be especially useful to psychoanalytic institutes and their candidates, where a careful and comprehensive assessment might increase probability of holding and effectively treating what are often tenuous control cases. Here, we assert that psychological testing can pinpoint ego strengths and deficits in ways that add incrementally to information gained during a trial or conversion period prior to recommending analysis proper. Such testing can also recommend against analysis and save the patient, candidate, and supervising analyst from the emotional strain and narcissistic injury of a failed treatment.

Moreover, inclusion of a module educating candidates about personality assessment in the psychoanalytic curriculum *Assessment* or *Analyzability* course can introduce candidates to the value of thorough personality assessment and also support recent discussion about the integration of psychoanalytic education with university faculty and academic institutions that develop the assessment methods (Kernberg, 2011).

Using psychological testing as an adjunct to psychoanalysis or psychoanalytic psychotherapy would also be analogous to medication consultation, where the consultant plays a role in the analytic treatment via direct service. Thus, it would appear that a psychoanalytically-informed approach to diagnostic psychological testing has a place in psychoanalysis as a complementary method of clinical assessment that can provide very useful information in particular circumstances to support decisions related to analyzability.

Selective use of psychological testing also dovetails with the American Psychoanalytic Association's (2006) practice guidelines for assessment of psychoanalytic patients including clarification of strengths and vulnerabilities related to motivation, self-observation, frustration tolerance, affect regulation, empathy, object relations, defenses, and reality testing. These areas are tied closely to psychological testing employing measures interpreted both nomothetically and idiographically.

Additionally, psychological testing, as advocated in this article, illuminates the configurational contexts that map the conditions under which a patient is more or less likely to regress, content themes tied to regression, capacity to self-observe during regression, and ability to recover with or without supportive intervention (Bram and Peebles, 2014; Peebles-Kleiger, 2002). We hope

that our explanations and case illustrations encourage clinical analysts to consider the possibility of referrals for testing, especially when stymied diagnostically, uncertain of treatment direction, or embroiled in therapeutic impasse.

ACKNOWLEDGMENTS

We thank Drs. Regina Koziyevskaya, Linda Helmig Bram, Steve Huprich, Jonathan Shedler, and Marci Bauman-Bork for their generous assistance.

REFERENCES

Ackerman, S. J., A. J. Clemence, R. Weatherill, & M. J. Hilsenroth. (1999), Use of the TAT in the assessment of *DSM*-IV Cluster B personality disorders. *J. Personality Assessment*, 73(3): 422–442.

Allison, J., S. J. Blatt, & C. N. Zimet. (1968), *The Interpretation of Psychological Tests*. New York, NY: Harper & Row.

American Psychiatric Association. (2013), *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. Washington, DC: American Psychiatric Press.

American Psychoanalytic Association. (2006), Practice Bulletin 7: Psychoanalytic Clinical Assessment. *American Psychoanalyst*, 40(3), Supplement: 1–8.

Andrews, G., M. Singh, & M. Bond. (1993), The Defense Style Questionnaire. *J. Nervous and Mental Disease*, 181: 246–256.

Appelbaum, A. H. (1989), Supportive therapy: A developmental view. In: *Supportive Therapy: A Psychodynamic Approach*, ed. L. H. Rockland. New York, NY: Basic Books, pp. 40–57.

Appelbaum, S. A. (1963), The masochistic character as self-saboteur (with special reference to psychological testing). *J. Projective Techniques*, 27(1): 35–46.

———. (1976), Objections to diagnosis and diagnostic testing diagnosed. *Bull. Menn. Clin.*, 40: 559–564.

———. (1977), *Anatomy of Change*. New York: Plenum Press.

Bachrach, H. M., & L. A. Leaff. (1978), Analyzability: A systematic review of the clinical and quantitative literature. *J. Amer. Psychoanal. Assn.*, 26: 881–920.

Bagby, R. M., J. D. A. Parker, R. T. Joffe, & R. Buis. (1994), Reconstruction and validation of the Depressive Experiences Questionnaire. *Assessment*, 1: 59–68.

Beck, A. T., R. A. Steer, & G. K. Brown. (1996), *Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Bell, M. D. (1995), *Bell Object Relations and Reality Testing Inventory (BORRTI) Manual*. Los Angeles, CA: Western Psychological.

———, B. Becker, & R. Billington. (1986), A scale for the assessment of object relations: Reliability, validity, and factorial invariance. *J. Clinic. Psych.*, 42: 733–741.

Betan, E., A. K. Heim, C. Zittel Conklin, & D. Westen. (2005), Countertransference phenomena and personality pathology in clinical practice: An empirical investigation. *Amer. J. Psych.*, 162(5): 890–898.

Blatt, S. J. (2004), *Experiences of Depression: Theoretical, Clinical, and Research Perspectives*. Washington, DC: American Psychological Association.

———. (2006), A personal odyssey. *J. Personality Assess.*, 87: 1–14.

———, J. P. D'Affiti, & D. M. Quinlan. (1976), Experiences of depression in normal young adults. *J. Abnorm. Psych.*, 85: 383–389.

Bornstein, R. F. (1998a), Implicit and self-attributed dependency needs in dependent and histrionic personality disorders. *J. Personal. Assess.*, 71: 1–14.

———. (1998b), Implicit and self-attributed dependency strivings: Differential relationships to laboratory and field measures of help-seeking. *J. Personal. and Social Psych.*, 75: 778–787.

———. (1999), Criterion validity of objective and projective dependency tests: A metaanalytic assessment of behavioral prediction. *Psychol. Assess.*, 11: 48–57.

———. (2002), A process dissociation approach to objective-projective test score interrelationships. *J. Personal. Assess.*, 78(1): 47–68.

———. (2007), Toward a process-based framework for classifying personality tests: Comment on Meyer and Kurtz (2006). *J. Personal. Assess.*, 89(2): 202–207.

———, & J. M. Masling. (2005), The Rorschach Oral Dependency Scale. In: *Scoring the Rorschach: Seven Validated Systems*, ed. R. F. Bornstein & J. M. Masling. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 135–158.

Bowlby, J. (1969), *Attachment and Loss: Vol. 1: Attachment*. New York, NY: Basic Books.

Bram, A. D. (2010), The relevance of the Rorschach and patient-examiner relationship in treatment planning and outcome assessment. *J. Personal. Assess.*, 92(2): 91–115.

———. (2013), Psychological testing and treatment implications: We can say more. *J. Personal. Assess.*, 95(4): 319–331.

———. (2014), Object relations, interpersonal functioning, and health in a nonclinical sample: Construct validation and norms for the TAT SCORS-G. *Psychoanal. Psych.*, 31: 314–342.

———, S. J. Gallant, & C. Segrin. (1999), A longitudinal investigation of object relations: Child-rearing antecedents, stability in adulthood, and construct validation. *J. Res. in Personal.*, 33: 159–188.

———, & M. J. Peebles. (2014), *Psychological Testing That Matters: Creating a Road Map Effective Treatment*. Washington, DC: APA Books.

Burns, B., & D. J. Viglione. (1996), The Rorschach Human Experience Variable, interpersonal relatedness, and object representation in nonpatients. *Psychol. Assess.*, 8: 92–99.

Butcher, J. N., W. G. Dahlstrom, J. R. Graham, A. Tellegen, & B. Kaemmer. (1989), *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for Administration and Scoring*. Minneapolis, MN: University of Minnesota Press.

———, C. L. Williams, J. R. Graham, R. P. Archer, A. Tellegen, Y. S. Ben-Porath, & B. Kaemmer. (1992), *Minnesota Multiphasic Personality Inventory-Adolescent Version (MMPI-A): Manual for Administration, Scoring and Interpretation*. Minneapolis, MN: University of Minnesota Press.

Cate, R., M. Khademi, P. Judd, & H. Miller. (2011), Deficits in mentalization as a risk factor in the future development of eating disorders. Poster presented at the conference of the American Psychoanalytic Association, New York City, January.

Clarkin, J. F., E. Caligor, B. Stern, & O. F. Kernberg. (2004), *Structured Interview of Personality Organization (STIPO)*. New York, NY: Weill Cornell Medical College.

Conklin, A. C., J. C. Malone, & J. T. Fowler. (2012), Mentalization and the Rorschach. *Rorschachiana*, 33(2): 189–213.

Conners, C. K., G. Sitarenios, J. D. A. Parker, & J. N. Epstein. (1998), The Revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *J. Abnorm. Child Psych.*, 26: 257–268.

Cooper, S. H., & D. Arnow. (1986), *The Rorschach Defense Scales* [Unpublished scoring manual]. Cambridge, MA: Department of Psychology, Cambridge Hospital and Harvard Medical School.

———, J. C. Perry, & M. O'Connell. (1991), The Rorschach Defense Scales II: Longitudinal perspectives. *J. Personal. Assess.*, 56(2): 191–201.

Cousineau, T. M., & J. Shedler. (2006), Predicting physical health: Implicit mental health measures versus self-report scales. *J. Nervous and Mental Disease*, 194(6): 427–432.

Cramer, P. (1991), *The Development of Defense Mechanisms: Theory, Research, and Assessment*. New York, NY: Springer-Verlag.

———. (2006), *Protecting the Self: Defense Mechanisms in Action*. New York, NY: Guilford.

Desmat, M., S. Vanheule, H. Goenvynck, P. Verhaeghe, J. Vogel, & S. Bogaerts. (2007), The Depressive Experiences Questionnaire: An inquiry into the different scoring procedures. *Europe. J. Psychol. Assess.*, 23(2): 89–98.

Eissler, K. R. (1953), The effect of the structure of the ego on psychoanalytic technique. *J. Amer. Psychoanal. Assn.*, 1: 104–143

Exner, J. E. (2003), *The Rorschach: A Comprehensive System Vol. 1: Basic Foundations and Principles of interpretation* (4th ed.). New York, NY: Wiley.

Fairbairn, W. (1952), *Psychoanalytic Studies of the Personality*. London: Tavistock.

Falkenstrom, F., O. A. Solbakken, C. Moller, B. Lech, R. Sandell, & R. Holmqvist. (2014), Reflective functioning, affect consciousness, and mindfulness: Are these different functions? *Psychoanal. Psych.*, 31: 26–40.

First, M. B., R. L. Spitzer, M. Gibbon, & J. B. W. Williams. (1996), *Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (CV)*. Washington, DC: American Psychiatric Press.

Flett, G. L., P. L. Hewitt, N. S. Endler, & R. M. Bagby. (1995), Conceptualization and assessment of personality factors in depression. *Europe. J. Personal.*, 9: 309–350.

Fonagy, P. (2001), *Attachment Theory and Psychoanalysis*. New York, NY: Other Press.

———, G. Gergely, E. L. Jurist, & M. Target. (2002), *Affect Regulation, Mentalization, and the Development of the Self*. New York, NY: Other Press.

———, M. Target, M. Steele, & H. Steele. (1998), *Reflective Functioning Manual: Version 5. For application to Adult Attachment Interviews*. London: University College.

Freud, S. (1904), Psycho-analytic procedure. In: *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 7), ed. J. Strachey. London: Hogarth Press, 1953, pp. 249–254.

Gabbard, G. O. (1989), Two subtypes of narcissistic personality disorder. *Bull. of the Menninger Clinic*, 53: 527–532.

———. (1995), Countertransference: The emerging common ground. *International Journal of Psychoanalysis*, 76: 475–485.

George, C., & M. L. West. (2012), *The Adult Attachment Projective Picture System: Attachment Theory and assessment in Adults*. New York, NY: Guilford.

Hansen, J. W. (2001), Review of empirical research that utilized the Bell Object Relations Inventory. Unpublished Doctoral Dissertation, Biola University, La Mirada, CA.

Heaton, R. K., G. J. Chelune, J. L. Talley, G. G. Kay, & G. Curtiss. (1993), *Wisconsin Card Sorting Test Manual: Revised and Expanded*. Odessa, FL: Psychological assessment Resources.

Hirschfeld, R. M. A., G. L. Klerman, H. G. Gough, J. Barrett, S. J. Korchin, & P. Chodoff. (1977), A Measure of Interpersonal Dependency. *J. Personal. Assess.*, 41: 610–618.

Horvath, A. O., A. C. Del Re, C. Fluckiger, & B. D. Symonds. (2011), Alliance in individual psychotherapy. *Psychother.: Theory, Research, Practice, Training*, 48(1): 9–16.

Huprich, S. K., ed. (2006), *Rorschach Assessment of the Personality Disorders*. Mahwah, NJ: Lawrence Erlbaum Associates.

———, & R. F. Bornstein. (2015), Behind closed doors: Sadomasochistic enactments and psychoanalytic research. *Psychoanal. Inq.*, 35(Suppl. 1): 185–195.

Jacobson, E. (1964), *Self and the Object World*. New York, NY: International Universities Press.

Karon, B. P. (2002), Analyzability or the ability to analyze? *Contemp. Psychoanal.*, 38: 121–140.

Kelly, F. D. (1997), *The Assessment of Object Relations Phenomena in Adolescents: TAT and Rorschach Measures*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kernberg, O. F. (1967), Borderline personality organization. *J. Amer Psychoanal. Assn.*, 15: 641–685.

———. (1976), *Object-Relations Theory and Clinical Psychoanalysis*. New York, NY: Jason Aronson.

———. (2011), Psychoanalysis and the University: A difficult relationship. *Internat. J. Psycho-Anal.*, 92: 609–622.

Kernhof, K. K., J. Kaufhold, & R. Grabhorn. (2008), Object relations and interpersonal problems in sexually abused female patients: An empirical study with the SCORS and IIP. *J. Personal. Assess.*, 90(1): 44–51.

Kleiger, J. H. (1999), *Disordered Thinking and the Rorschach: Theory, Research, and Differential Diagnosis*. Hillsdale, NJ: The Analytic Press.

Klein, M., & J. Riviere. (1964) *Love, Hate, and Reparation*. New York, NY: W.W. Norton & Company.

Kohut, H. (1972), Thoughts on narcissism and narcissistic rage. *Psychoanal. Study of the Child*, 27, 360–400.

———. (1984), *How Does Analysis Cure?* Chicago, IL: University of Chicago Press.

Lensenweger, M. F., J. F. Clarkin, O. F. Kernberg, & P. A. Foelsch. (2001), The Inventory of Personality Organization: Psychometric properties, factorial composition, and criterion relations with affect, aggressive dyscontrol, psychosis proneness, and self domains in a nonclinical sample. *Psycholog. Assess.*, 13: 577–591.

Lerner, P. (1998), *Psychoanalytic Perspectives on the Rorschach*. Hillsdale, NJ: The Analytic Press.

Mahler, M., F. Pine, & A. Bergman. (1975), *The Psychological Birth of the Human Infant*. New York, NY: Basic Books.

Masling, J., L. Rabie, & S. H. Blondheim. (1967), Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *J. Consulting Psych.*, 31(3): 233–239.

Mayman, M. (1968), Early memories and character structure. *J. Projective Tech. and Personal. Assess.*, 32: 303–316.

McClelland, D. C., R. Koestner, & J. Weinberger. (1989), How do self-attributed and implicit motives differ? *Psycholog. Rev.*, 96(4): 690–702.

McWilliams, N. (1999), *Psychoanalytic Case Formulation*. New York, NY: Guilford.

———. (2011), *Psychoanalytic Diagnosis*, 2nd edition. New York, NY: Guilford.

Meyer, G. J., S. E. Finn, L. D. Eyde, G. G. Kay, K. L. Moreland, R. R. Dies, . . . G. M. Reed. (2001), Psychological testing and assessment: A review of evidence and issues. *Amer. Psych.*, 56(2): 128–165.

————, W. Hsiao, D. J. Viglione, J. L. Mihura, & J. M. Abraham. (2013), Rorschach scores in applied clinical practice: A survery of perceived validity by experienced clinicians. *J. Personal. Assess.*, 95: 351–365.

————, D. J. Viglione, J. L. Mihura, R. E. Erard, & P. Erdberg. (2011), *Rorschach Performance Assessment System: Administration, Coding, Interpretation, and Technical Manual.* Toledo, OH: Rorschach Performance Assessment System, LLC.

Mihura, J. L., G. J. Meyer, N. Dumitrascu, & G. Bombel. (2013), The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the comprehensive system. *Psycholog. Bull.*, 139: 548–605.

Monroe, J. M., M. Diener, J. C. Fowler, J. E. Sexton, & M. J. Hilsenroth. (2013), Criterion validity of the Rorschach Mutuality of Autonomy (MOA) scale: A meta-analytic review. *Psychoanal. Psych.*, 30(4): 535–566.

Morey, L. C. (1991), *Personality Assessment Inventory Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Murray, H. (1943), *Thematic Apperception Test Manual.* Cambridge, MA: Harvard University Press.

PDM Task Force. (2006), *Psychodynamic Diagnostic Manual.* Silver Springs, MD: Alliance of Psychoanalytic Organizations.

Peebles, M. J. (2012), *Beginnings: The Art and Science of Planning Psychotherapy*, 2nd edition. New York, NY: Routledge.

Peebles-Kleiger, M. J. (2002), Elaboration of some sequence analysis strategies: Examples and guidelines for level of confidence. *J. Personal. Assess.*, 79(1): 19–38.

————, L. Horwitz, J. H. Kleiger, & R. M. Waugaman. (2006), Psychological testing and analyzability: Breathing new life into an old issue. *Psychoanal. Psych.*, 23: 5040–526.

Piaget, J. (1937), *The Construction of Reality in the Child*, trans. M. Cook. New York: Basic Books, 1954.

Rapaport, D., M. Gill, & R. Schafer. (1968), *Diagnostic Psychological Testing* (rev. ed.). New York: International Universities Press.

Rudolph, A., M. Schroder, & A. Schutz. (2006), Ein impliziter Assoziationstest zur Erfassung Von Selbstwertschaätzung [An implicit association test measuring self-esteem]. In: *Theorie und Praxis Objektiver Personlichkeitstests*, ed. T. M. Ortner, R. T. Proyer, & K. D. Kubinger. Bern, Switzerland: Verlag Hans Huber, pp. 153–163.

Schafer, R. (1954), *Psychoanalytic Interpretation of Rorschach Testing.* New York, NY: Grune & Stratton.

————. (1958), How was this story told? *J. Projective Tech.*, 22: 181–210.

————. (1967), *Projective Testing and Psychoanalysis.* New York, NY: International Universities Press.

————. (2006), My life in testing. *J. Personal. Assess.*, 86(3): 235–241.

Schlesinger, H. J. (1995), The process of interpretation and the moment of change. *J. Amer. Psychoanal. Assn*, 43: 633–688.

Schroder-Abe, M., A. Rudolph, & A. Schutz. (2007), High implicit self-esteem is not necessarily advantageous: Discrepancies between explicit and implicit self-esteem and their relationship with anger expression and psychological health. *Europe. J. Personal.*, 21: 319–339.

Schutz, A., & I. Sellin. (2006), *Die multidimensionale Selbstwertskala (MSWS) [The Multidimensional self-esteem scale].* Gottingen, Germany: Hogrefe.

Shectman, F., & M. K. Harty. (1986), Treatment implications of object relationships as they unfold during the diagnostic interaction. In: *Assessing Object Relations Phenomena*, ed. M. Kissen. Madison, CT: International Universities Press, pp. 279–303.

Shedler, J., M. Mayman, & M. Manis. (1993), The illusion of mental health. *Amer. Psychol.*, 48(11): 1117–1131.

————, & D. Westen. (2007), The Shedler-Westen Assessment Procedure: Making personality diagnosis clinically meaningful. *J. Personal. Assess.*, 89(1): 41–55.

Slavin, J. M., M. S. Stein, J. H. Pinsker-Aspen, & M. J. Hilsenroth. (2007), Early memories from outpatients with and without a history of childhood sexual abuse. *J. Loss and Trauma*, 12: 435–451.

Smits, J. M., R. Vermote, L. Claes, & H. Vertommen. (2009), Inventory of Personality Organization-Revised: Constuction of an abridged version. *European Journal of Psychological Assessment*, 25: 223–230.

Stein, H., N. J. Jacobs, K. S. Ferguson, J. G. Allen, & P. Fonagy. (1998), What do adult attachment scales measure? *Bull. Menn. Clin.*, 62(1): 33–82.

Stein, M. B., M. J. Hilsenroth, J. Slavin-Mumford, & J. Pinsker. (2011), *Social Cognition and Object Relations Scale: Global Rating Method (SCORS-G;* 4th edition) [Unpublished manuscript]. Boston, MA: Massachusetts General Hospital and Harvard Medical School.

Stone, L. (1954), The widening scope of indications for psychoanalysis. *J. Amer. Psychoanal. Assn.*, 2: 567–594.

Stricker, G., & B. Healey. (1990), Projective assessment of object relations: A review of the empirical literature. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 2(3): 219–230.

Symons, D. K., C. C. Peterson, V. Slaughter, J. Roche, & E. Doyle. (2005), Theory of mind and mental state discourse during book reading and story-telling tasks. *Brit. J. Develop. Psych.*, 23: 81–102.

Tuber, S. B. (2012), *Understanding Personality Through Projective Testing*. Lanham, MD: Jason Aronson.

Urist, J. (1977), The Rorschach Test and the assessment of object relations. *J. Personal. Assess.*, 41:, 3–9.

Wallerstein, R. (1986), *Forty-Two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy*. New York, NY: Guilford.

Werner, H. (1957), The concept of development from a comparative and organismic view. In: *The Concept of Development*, ed. D. B. Harris. Minneapolis: University of Minnesota Press, pp. 125–148.

Westen, D. (1995), *Social Cognition and Object Relations Scale: Q-Sort for Projective Stories (SCORS-Q)*. Cambridge, MA: Department of Psychiatry, Cambridge Hospital and Harvard Medical School.

———. (2004), *Clinical Diagnostic Interview*. Atlanta, GA: Departments of Psychology and Psychiatry, Emory University. Retrieved from http://www.swapassessment.org/supporting-documents/l

———. (2008), *Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV)*. San Antonio, TX: Psychological Corporation.

———, N. Lohr, K. Silk, & K. Kerber. (1985), *Measuring Object Relations and Social Cognition Using the TAT: Scoring Manual*. University of Michigan.

———, & S. Muderrisoglu. (2003), Reliability and validity of personality disorder assessment using a systematic clinical interview: Evaluating an alternative to structured interviews. *J. Personal. Disord.*, 17: 350–368.

Winnicott, D. W. (1965), *The Maturational Processes and the Facilitating Environment: Studies in the Theory of Emotional Development*. New York, NY: International Universities Press.

Worchel, S., J. Cooper, & G. R. Goethals. (1991), *Understanding Social Psychology*, 5th edition. Homewood, IL: Dorsey Press.

World Health Organization. (in press), *International Classification of Diseases and Related Health Problems*, 11th edition. Geneva, Switzerland: Author.

Yalof, J. (2006), Case illustration of a boy with Nonverbal Learning Disorder and Asperger's features. Neuropsychological and personality assessment. *J. Personal. Assess.*, 87: 15–34.

———. (2015), Teaching psychoanalytic concepts in the university setting: Issues, challenges, and promises. *Psychoanal. Inq.*, 35(Suppl. 1): 124–134.

———, & D. Rosenstein. (2014), Psychoanalytic interpretation of superego functioning following CS readministration procedures: Case illustration. *J. Personal. Assess.*, 96(2): 192–203, doi:10.1080/00223891.2013.836528

Anthony D. Bram, Ph.D.
363 Massachusetts Avenue, LL#1
Lexington, MA 02420
Anthony_Bram@hms.harvard.edu

Neuron

Fractionating Human Intelligence

Adam Hampshire,[1,*] Roger R. Highfield,[2] Beth L. Parkin,[1] and Adrian M. Owen[1]
[1]The Brain and Mind Institute, The Natural Sciences Centre, Department of Psychology, The University of Western Ontario, London ON, N6A 5B7, Canada
[2]Science Museum, Exhibition Road, London SW72DD, UK
*Correspondence: ahampshi@uwo.ca
http://dx.doi.org/10.1016/j.neuron.2012.06.022

SUMMARY

What makes one person more intellectually able than another? Can the entire distribution of human intelligence be accounted for by just one general factor? Is intelligence supported by a single neural system? Here, we provide a perspective on human intelligence that takes into account how general abilities or "factors" reflect the functional organization of the brain. By comparing factor models of individual differences in performance with factor models of brain functional organization, we demonstrate that different components of intelligence have their analogs in distinct brain networks. Using simulations based on neuroimaging data, we show that the higher-order factor "g" is accounted for by cognitive tasks corecruiting multiple networks. Finally, we confirm the independence of these components of intelligence by dissociating them using questionnaire variables. We propose that intelligence is an emergent property of anatomically distinct cognitive systems, each of which has its own capacity.

INTRODUCTION

Few topics in psychology are as old or as controversial as the study of human intelligence. In 1904, Charles Spearman famously observed that performance was correlated across a spectrum of seemingly unrelated tasks (Spearman, 1904). He proposed that a dominant general factor "g" accounts for correlations in performance between all cognitive tasks, with residual differences across tasks reflecting task-specific factors. More controversially, on the basis of subsequent attempts to measure "g" using tests that generate an intelligence quotient (IQ), it has been suggested that population variables including gender (Irwing and Lynn, 2005; Lynn, 1999), class (Burt, 1959, 1961; McManus, 2004), and race (Rushton and Jensen, 2005) correlate with "g" and, by extension, with one's genetically predetermined potential. It remains unclear, however, whether population differences in intelligence test scores are driven by heritable factors or by other correlated demographic variables such as socioeconomic status, education level, and motivation (Gould, 1981; Horn and Cattell, 1966). More relevantly, it is questionable whether they relate to a unitary intelligence factor, as opposed to a bias in testing paradigms toward particular components of a more complex intelligence construct (Gould, 1981; Horn and Cattell, 1966; Mackintosh, 1998). Indeed, over the past 100 years, there has been much debate over whether general intelligence is unitary or composed of multiple factors (Carroll, 1993; Cattell, 1949; Cattell and Horn, 1978; Johnson and Bouchard, 2005). This debate is driven by the observation that test measures tend to form distinctive clusters. When combined with the intractability of developing tests that measure individual cognitive processes, it is likely that a more complex set of factors contribute to correlations in performance (Carroll, 1993).

Defining the biological basis of these factors remains a challenge, however, due in part to the limitations of behavioral factor analyses. More specifically, behavioral factor analyses do not provide an unambiguous model of the underlying cognitive architecture, as the factors themselves are inaccessible, being measured indirectly by estimating linear components from correlations between the performance measures of different tests. Thus, for a given set of behavioral correlations, there are many factor solutions of varying degrees of complexity, all of which are equally able to account for the data. This ambiguity is typically resolved by selecting a simple and interpretable factor solution. However, interpretability does not necessarily equate to biological reality. Furthermore, the accuracy of any factor model depends on the collection of a large number of population measures. Consequently, the classical approach to intelligence testing is hampered by the logistical requirements of pen and paper testing. It would appear, therefore, that the classical approach to behavioral factor analysis is near the limit of its resolution.

Neuroimaging has the potential to provide additional constraint to behavioral factor models by leveraging the spatial segregation of functional brain networks. For example, if one homogeneous system supports all intelligence processes, then a common network of brain regions should be recruited whenever difficulty increases across all cognitive tasks, regardless of the exact stimulus, response, or cognitive process that is manipulated. Conversely, if intelligence is supported by multiple specialized systems, anatomically distinct brain networks should be recruited when tasks that load on distinct intelligence factors are undertaken. On the surface, neuroimaging results accord well with the former account. Thus, a common set of frontal and parietal brain regions is rendered when peak activation coordinates from a broad range of tasks that parametrically modulate difficulty are smoothed and averaged (Duncan and Owen, 2000). The same set of multiple demand (MD) regions is activated during tasks that load on "g" (Duncan, 2005; Jung

and Haier, 2007), while the level of activation within frontoparietal cortex correlates with individuals differences in IQ score (Gray et al., 2003). Critically, after brain damage, the size of the lesion within, but not outside of, MD cortex is correlated with the estimated drop in IQ (Woolgar et al., 2010). However, these results should not necessarily be equated with a proof that intelligence is unitary. More specifically, if intelligence is formed from multiple cognitive systems and one looks for brain responses during tasks that weigh most heavily on the "g" factor, one will most likely corecruit all of those functionally distinct systems. Similarly, by rendering brain activation based on many task demands, one will have the statistical power to render the networks that are most commonly recruited, even if they are not always corecruited. Indeed, there is mounting evidence demonstrating that different MD regions respond when distinct cognitive demands are manipulated (Corbetta and Shulman, 2002; D'Esposito et al., 1999; Hampshire and Owen, 2006; Hampshire et al., 2008, 2011; Koechlin et al., 2003; Owen et al., 1996; Petrides, 2005). However, such a vast array of highly specific functional dissociations have been proposed in the neuroimaging literature as a whole that they often lack credibility, as they fail to account for the broader involvement of the same brain regions in other aspects of cognition (Duncan and Owen, 2000; Hampshire et al., 2010). The question remains, therefore, whether intelligence is supported by one or multiple systems, and if the latter is the case, which cognitive processes those systems can most broadly be described as supporting. Furthermore, even if multiple functionally distinct brain networks contribute to intelligence, it is unknown whether the capacities of those networks are independent or are related to the same set of diffuse biological factors that modulate general neural efficiency. It is unclear, therefore, whether the pattern of individual differences in intelligence reflects the functional organization of the brain.

Here, we address the question of whether human intelligence is best conceived of as an emergent property of functionally distinct brain networks using factor analyses of brain imaging, behavioral, and simulated data. First, we break MD cortex down into its constituent functional networks by factor analyzing regional activation levels during the performance of 12 challenging cognitive tasks. Then, we build a model, based on the extent to which the different functional networks are recruited during the performance of those 12 tasks, and determine how well that model accounts for cross-task correlations in performance in a large (n = 44,600) population sample. Factor solutions, generated from brain imaging and behavioral data, are compared directly, to answer the question of whether the same set of cognitive entities is evident in the functional organization of the brain and in individual differences in performance. Simulations, based on the imaging data, are used to determine the extent to which correlations between first-order behavioral components are predicted by cognitive tasks recruiting multiple functional brain networks, and the extent to which those correlations may be accounted for by a spatially diffuse general factor. Finally, we examine whether the behavioral components of intelligence show a degree of independence, as evidenced by dissociable correlations with the types of questionnaire variable that "g" has historically been associated with.

RESULTS AND DISCUSSION

Identifying Functional Networks within MD Cortex

Sixteen healthy young participants undertook the cognitive battery in the MRI scanner. The cognitive battery consisted of 12 tasks, which, based on well-established paradigms from the neuropsychology literature, measured a range of the types of planning, reasoning, attentional, and working memory skills that are considered akin to general intelligence (see Supplemental Experimental Procedures available online). The activation level of each voxel within MD cortex was calculated separately for each task relative to a resting baseline using general linear modeling (see Supplemental Experimental Procedures) and the resultant values were averaged across participants to remove between-subject variability in activation—for example, due to individual differences in regional signal intensity.

The question of how many functionally distinct networks were apparent within MD cortex was addressed using exploratory factor analysis. Voxels within MD cortex (Figure 1A) were transformed into 12 vectors, one for each task, and these were examined using principal components analysis (PCA), a factor analysis technique that extracts orthogonal linear components from the 12-by-12 matrix of task-task bivariate correlations. The results revealed two "significant" principal components, each of which explained more variability in brain activation than was contributed by any one task. These components accounted for ~90% of the total variance in task-related activation across MD cortex (Table S1). After orthogonal rotation with the Varimax algorithm, the strengths of the task-component loadings were highly variable and easily comprehensible (Table 1 and Figure 1B). Specifically, all of the tasks in which information had to be actively maintained in short-term memory, for example, spatial working memory, digit span, and visuospatial working memory, loaded heavily on one component (MDwm). Conversely, all of the tasks in which information had to be transformed in mind according to logical rules, for example, deductive reasoning, grammatical reasoning, spatial rotations, and color-word remapping, loaded heavily on the other component (MDr). When factor scores were generated at each voxel using regression and projected back onto the brain, two clearly defined functional networks were rendered (Figure 1D). Thus, the insula/frontal operculum (IFO), the superior frontal sulcus (SFS), and the ventral portion of the anterior cingulate cortex/ presupplementary motor area (ACC/preSMA) had greater MDwm component scores, whereas the inferior frontal sulcus (IFS), inferior parietal cortex (IPC), and the dorsal portion of the ACC/preSMA had greater MDr component scores. When the PCA was rerun with spherical regions of interest (ROIs) centered on each MD subregion, with radii that varied from 10 to 25 mm in 5 mm steps and excluding voxels that were on average deactivated, the task loadings correlated with those from the MD mask at $r > 0.95$ for both components and at all radii. Thus, the PCA solution was robust against variations in the extent of the ROIs. When data from the whole brain were analyzed using the same method, three significant components were generated, the first two of which correlated with those from the MD cortex analysis (MDr $r = 0.76$, MDwm $r = 0.83$), demonstrating that these were the most prominent active-state networks in the brain. The factor solution was also reliable at

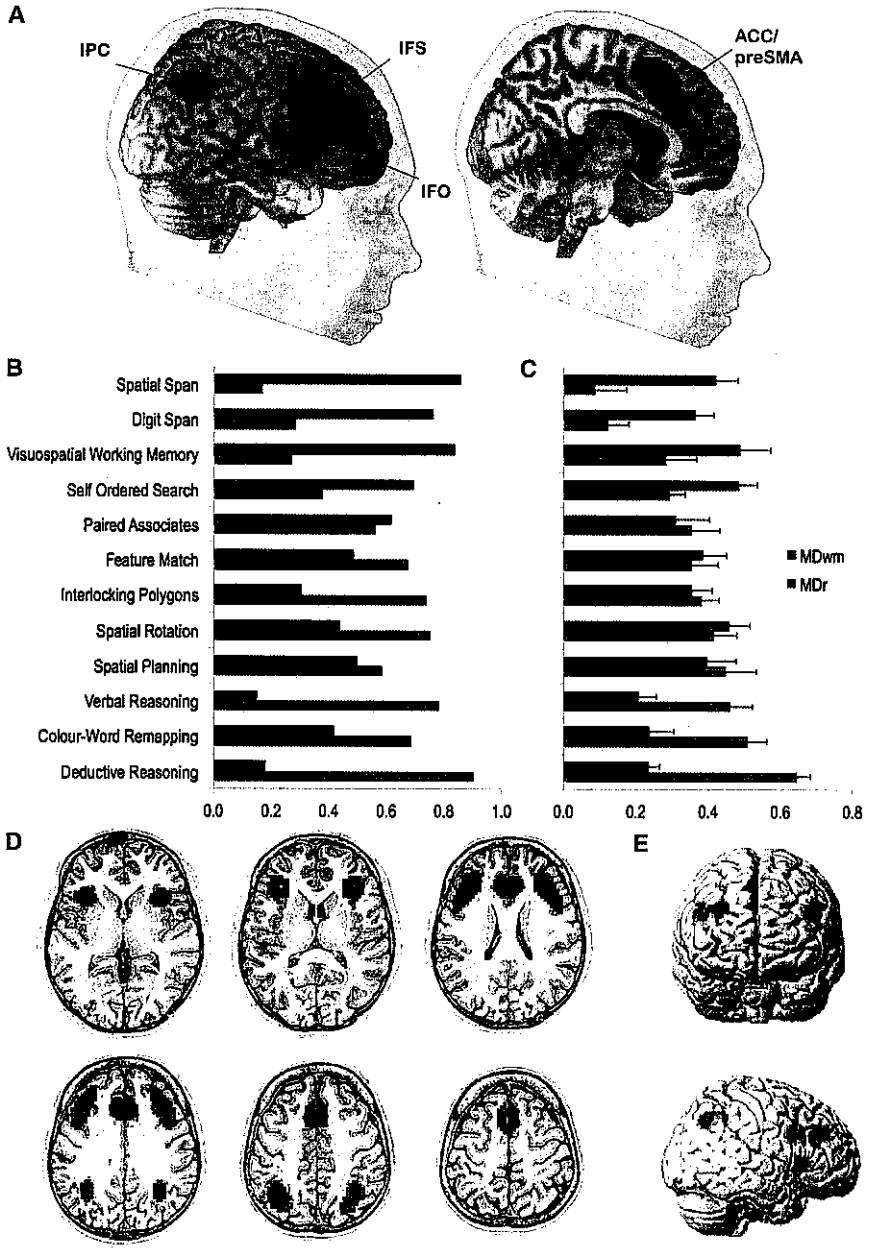

Figure 1. Factor Analyzing Functional Brain Imaging Data from within Multiple Demand Cortex

(A) The MD cortex ROIs.

(B) PCA of the average activation patterns within MD cortex for each task (x axis reports task-component loading).

(C) PCA with each individual's data included as separate columns (error bars report SEM).

(D) Component scores from the analysis of MD task-related activations averaged across individuals. Voxels that loaded more heavily on the MDwm component are displayed in red. Voxels that loaded more heavily on the MDr network are displayed in blue.

(E) T contrasts of component scores against zero from the PCA with individual data concatenated into 12 columns (FDR corrected at $p < 0.05$ for all MD voxels).

the individual subject level. Rerunning the same PCA on each individual's data generated solutions with two significant components in 13/16 cases. There was one three-component solution and two four-component solutions. Rerunning the two-component PCA with each individual's data set included as 12 separate columns (an approach that did not constrain the same task to load on the same component across participants) demonstrated that the pattern of task-component loadings was also highly reliable at the individual subject level (Figure 1C). In order to test the reliability of the functional networks across participants, the data were concatenated instead of averaged into 12 columns (an approach that does not constrain the same voxels to load on the same components across individuals), and component scores were estimated at each voxel and projected back into two sets of 16 brain maps. When t contrasts were calculated against zero at the group level, the same MDwm and MDr functional networks were rendered (Figure 1E).

While the PCA works well to identify the number of significant components, a potential weakness for this method is that the unrotated task-component loadings are liable to be formed from mixtures of the underlying factors and are heavily biased toward the component that is extracted first. This weakness necessitates the application of rotation to the task-component matrix; however, rotation is not perfect, as it identifies the task-component loadings that fit an arbitrary set of criteria designed to generate the simplest and most interpretable solution. To deal with this potential issue, the task-functional network loadings were recalculated using independent component analysis (ICA), an analysis technique that exploits the more powerful properties of statistical independence to extract the sources from mixed signals. Here, we used ICA to extract two spatially distinct functional brain networks using gradient ascent toward maximum entropy (code adapted from Stone and Porrill, 1999). The resultant components were broadly similar, although not identical, to those from the PCA (Table 1). More specifically, all tasks loaded positively on both independent brain networks but to highly varied extents, with the short-term memory tasks loading heavily on one component and the tasks that involved transforming information according to logical rules loading heavily on the other. Based on these results, it is reasonable to conclude that MD cortex is formed from at least two functional networks, with all 12 cognitive tasks recruiting both networks but to highly variable extents.

Table 1. PCA and ICA of Activation Levels in 2,275 MD Voxels during the Performance of 12 Cognitive Tasks

	PCA		ICA	
	MDr	MDwm	MDr	MDwm
Self-ordered search	0.38	0.69	1.45	3.26
Visuospatial working memory	0.27	0.84	1.24	2.68
Spatial span	0.17	0.86	0.51	2.23
Digit span	0.28	0.76	0.76	2.20
Paired associates	0.56	0.62	1.90	1.97
Spatial planning	0.58	0.50	2.43	2.74
Feature match	0.68	0.49	2.00	0.88
Interlocking polygons	0.74	0.31	2.11	0.61
Verbal reasoning	0.78	0.15	2.62	0.60
Spatial rotation	0.75	0.44	2.86	1.88
Color-word remapping	0.69	0.42	3.07	0.95
Deductive reasoning	0.90	0.18	3.98	0.19
PCA/ICA correlation MDr	r = 0.92			
PCA/ICA correlation MDwm	r = 0.81			

Table 2. Task-Component Loadings from the PCA of Internet Data with Orthogonal Rotation

	1 (STM)	2 (Reasoning)	3 (Verbal)
Spatial span	0.69	0.22	
Visuospatial working memory	0.69	0.21	
Self-ordered search	0.62	0.16	0.16
Paired associates	0.58		0.25
Spatial planning	0.41	0.45	
Spatial rotation	0.14	0.66	
Feature match	0.15	0.57	0.22
Interlocking polygons		0.54	0.3
Deductive reasoning	0.19	0.52	−0.14
Digit span	0.26	−0.2	0.71
Verbal reasoning		0.33	0.66
Color-word remapping	0.22	0.35	0.51

The Relationship between the Functional Organization of MD Cortex and Individual Differences in Intelligence: Permutation Modeling

A critical question is whether the loadings of the tasks on the MDwm and MDr functional brain networks form a good predictor of the pattern of cross-task correlations in performance observed in the general population. That is, does the same set of cognitive entities underlay the large-scale functional organization of the brain and individual differences in performance? It is important to note that factor analyses typically require many measures. In the case of the spatial factor analyses reported above, measures were taken from 2,275 spatially distinct "voxels" within MD cortex. In the case of the behavioral analyses, we used scores from ~110,000 participants who logged in to undertake Internet-optimized variants of the same 12 tasks. Of these, ~60,000 completed all 12 tasks and a post task questionnaire. After case-wise removal of extreme outliers, null values, nonsense questionnaire responses, and exclusion of participants above the age of 70 and below the age of 12, exactly 44,600 data sets, each composed of 12 standardized task scores, were included in the analysis (see Experimental Procedures).

The loadings of the tasks on the MDwm and MDr networks from the ICA were formed into two vectors. These were regressed onto each individual's set of 12 standardized task scores with no constant term. When each individual's MDwm and MDr beta weights (representing component scores) were varied in this manner, they centered close to zero, showed no positive correlation (MDwm mean beta = 0.05 ± 1.78; MDr mean beta = 0.11 ± 2.92; MDwm-MDr correlation r = −0.20), and, importantly, accounted for 34.3% of the total variance in performance scores. For comparison, the first two principal components of the behavioral data accounted for 36.6% of the variance. Thus, the model based on the brain imaging data captured close to the maximum amount of variance that could be accounted for by the two best-fitting orthogonal linear components. The average test-retest reliability of the 12 tasks, collected in an earlier Internet cohort (Table S2), was 68%. Consequently, the imaging ICA model predicted >50% of the reliable variance in performance. The statistical significance of this fit was tested against 1,000 permutations, in which the MDwm and MDr vectors were randomly rearranged both within and across vector prior to regression. The original vectors formed a better fit than the permuted vectors in 100% of cases, demonstrating that the brain imaging model was a significant predictor of the performance data relative to models with the same fine-grained values and the same level of complexity. Two further sets of permutation tests were carried out in which one vector was held constant and the other randomly permuted 1,000 times. When the MDwm vector was permuted, the original vectors formed a better fit in 100% of cases. When the MDr vector was permuted, the original vectors formed a better fit in 99.3% of cases. Thus, both the MDwm and the MDr vectors were significant predictors of individual differences in behavioral performance.

The Relationship between the Functional Organization of MD Cortex and Individual Differences in Intelligence: Similarity of Factor Solutions

Exploratory factor analysis was carried out on the behavioral data using PCA. There were three significant behavioral components that each accounted for more variance than was contributed by any one test (Table S3) and that together accounted for 45% of the total variance. After orthogonal rotation with the Varimax algorithm, the first two components showed a marked similarity to the loadings of the tasks on the MDwm and MDr networks (Table 2). Thus, the first component (STM) included all of the tasks in which information was held actively on line in short-term memory, whereas the second component (reasoning) included all of the tasks in which information was transformed in mind according to logical rules. Correlation analyses between the task to functional brain network loadings and the task to behavioral component loadings confirmed that the two approaches generated broadly similar solutions (STM-MDwm

Cornell HR Review

1-26-2013

Personality Tests in Employment Selection: Use With Caution

H. Beau Baez
Charlotte School of Law

Personality Tests in Employment Selection: Use With Caution

Abstract
[Excerpt] Many employers utilize personality tests in the employment selection process to identify people who have more than just the knowledge and skills necessary to be successful in their jobs.[1] If anecdotes are to be believed—Dilbert must be getting at something or the cartoon strip would not be so popular—the work place is full of people whose personalities are a mismatch for the positions they hold. Psychology has the ability to measure personality and emotional intelligence ("EQ"), which can provide employers with data to use in the selection process. "Personality refers to an individual's unique constellation of consistent behavioral traits"[2] and "emotional intelligence consists of the ability to perceive and express emotion, assimilate emotion in thought, understand and reason with emotion, and regulate emotion."[3] By using a scientific approach in hiring, employers can increase their number of successful employees.

Keywords
HR Review, Human Resources, employment selection, personality tests

Disciplines
Human Resources Management | Labor Relations

Comments
Suggested Citation:
Baez H. (2013, January 26). Personality tests in employment selection: Use with caution. *Cornell HR Review*. Retrieved [insert date] from Cornell University, ILR School site: http://digitalcommons.ilr.cornell.edu/chrr/59

CORNELL HR REVIEW

PERSONALITY TESTS IN EMPLOYMENT SELECTION: USE WITH CAUTION

H. Beau Beaz III

Introduction

Many employers utilize personality tests in the employment selection process to identify people who have more than just the knowledge and skills necessary to be successful in their jobs.[1] If anecdotes are to be believed—Dilbert must be getting at something or the cartoon strip would not be so popular—the work place is full of people whose personalities are a mismatch for the positions they hold. Psychology has the ability to measure personality and emotional intelligence ("EQ"), which can provide employers with data to use in the selection process. "Personality refers to an individual's unique constellation of consistent behavioral traits"[2] and "emotional intelligence consists of the ability to perceive and express emotion, assimilate emotion in thought, understand and reason with emotion, and regulate emotion."[3] By using a scientific approach in hiring, employers can increase their number of successful employees.

Personality & Emotional Intelligence

The link between personality and emotional intelligence to job performance is compelling.[4] Though there is strong evidence that cognitive measurement tools are good predictors of job success, one important reason that they are not perfect predictors is that human personality is an important factor in job success.[5] But not all are convinced that assessing workers' cognitive abilities is worthwhile. Annie Murphy Paul, a former senior editor for *Psychology Today* magazine, attacked the $400 million a year testing industry, comparing personality tests to phrenology—a popular and discredited 19th century personality instrument that measured mental traits by examining the 27 bumps on a person's head.[6] With over 2,500 personality and emotional intelligence instruments on the market, Ms. Murphy is likely correct that some of these are ineffective.[7] Discernment is the solution.

Personality

Personality is "the sum total of ways in which an individual reacts to and interacts with others … [and] we most often describe it in terms of the measurable traits a person exhibits."[8] One of the best supported models for measuring personality involves the "Big Five Model," with its five basic dimensions that capture most of the variation in human personality.[9] The traits include neuroticism/emotional stability,[10] extraversion,[11] openness to experience,[12] agreeableness,[13] and conscientiousness.[14] These five job traits

are connected to job performance and are predictors of certain outcomes: "avoiding counterproductive behavior, reducing turnover and absenteeism, exhibiting more teamwork and leadership, providing more effective customer service, contributing more citizenship behavior, influencing job satisfaction and commitment to the firm, and enhancing safety."[15]

There are several tests that measure the Big Five personality dimensions, but the two most popular are the NEO-Personality Inventory and the Personality Characteristics Inventory ("PCI").[16] The PCI is comprised of 150 multiple-choice questions and asks questions such as "I tend not to say what I think about things" (i.e., testing extraversion) or "I approach most of my work steadily and persistently" (i.e., testing conscientiousness).[17] The first Big Five personality test developed for the business community was the Hogan Personality Inventory ("HPI"), with its focus on normal personality rather than abnormal personality.[18] A 2003 meta-analytic review of 43 studies found that the HPI is an effective predictor of job performance for many different jobs, including customer service representatives, hospital administrators, bus drivers, department managers, and police officers.[19]

Personality Test Criticism

There is some debate in the industrial/organizational ("IO") psychology field as to whether personality measures should be used in employee selection.[20] Many believe that personality tests used for employee selection are not valid, and in any event, can be faked.[21] The earliest personality tests go back at least to 1919, at the dawn of IO psychology.[22] In one article that reviewed 113 personality selection tool studies conducted from 1919 to 1952, personality was found to correlate to job success at levels similar to more recent studies.[23] For studies published from 1952 to 1963, one paper noted that the studies indicated that personality had some predictive power, but not at a level that personality should be used for employee selection.[24] This same article concluded that

> "there is no generalizable evidence that personality measures can be recommended as good or practical tools for employee selection.... The best that can be said is that in *some* situations, for *some* purposes, *some* personality measures can offer helpful predictions. But there is nothing in this summary to indicate in advance which measure should be used in which situation or for which purposes. In short, it must be concluded (as always) that the validity of any personality measure must be specifically and competently determined for the specific situation in which it is to be used and for the specific purpose or criterion within that situation.... It seems clear that the only acceptable reason for using personality measures as instruments of decision is found only after doing considerable research with the measure in the specific situation and for the specific purpose for which it is to be used.[25]"

A 2010 review of the academic literature found correlations between personality and job success to fall in the .03 to .15 range, which the authors note is "close to zero."[26] To put

these correlations in perspective, personality tests used in employee selection account for approximately 5% of an employee's job success while the other 95% of their performance is unaccounted for by personality.[27] Interestingly, the .15 correlation is almost identical to what was noted in the 1960's, meaning there has been no measurable change in the data for the 50 years.

One possibility for the relatively low correlation rates is that the data has not been interpreted properly. A 2011 study has found evidence for a curvilinear relationship between personality traits and job performance, while all the earlier studies assumed a linear relationship.[28] This suggests that for complex jobs, high personality scores may correlate better to ultimate job success.[29]

Emotional Intelligence

As the name implies, emotional intelligence ("EQ") is not a personality trait but a type of intelligence. Beginning in the 20th century, society has viewed intelligence almost exclusively through the lens of intelligence quotient ("IQ") tests.[30] IQ tests have the advantage of being very reliable, but they are limited in that they measure abstract reasoning and verbal fluency.[31] In 1990 Peter Salovey and John Mayer proposed an additional intelligence: emotional intelligence.[32] Emotional intelligence is comprised of four components: First, people need to be able to accurately perceive emotions in themselves and others and have the ability to express their own emotions effectively. Second, people need to be aware of how their emotions shape their thinking, decisions, and coping mechanisms. Third, people need to be able to understand and analyze their emotions, which may often be complex and contradictory. Fourth, people need to be able to regulate their emotions so that they can dampen negative emotions and make effective use of positive emotions.[33]

It is important to note that if EQ is, in fact, a type of intelligence, it really cannot be changed very much—just like a person's IQ remains relatively constant throughout their lifetime.

The marketplace is beginning to recognize the importance of EQ. One survey indicated that 60% of employers would not hire a high IQ candidate with a low EQ.[34]

When asked why emotional intelligence is more important than high IQ, employers said that employees with high EQ (in order of importance):

- Are more likely to stay calm under pressure
- Know how to resolve conflict effectively
- Are empathetic to their team members and react accordingly
- Lead by example
- Tend to make more thoughtful business decisions[35]

When these same employers were asked to identify specific behaviors and qualities that demonstrate EQ, they responded that employees who demonstrate high EQ:

- Admit and learn from their mistakes
- Can keep their emotions in check and have thoughtful discussions on tough issues
- Listen as much, or more than, they talk
- Take criticism well
- Show grace under pressure[36]

The opinions given by the surveyed employers are also echoed in academic literature on the subject. Research indicates that emotional intelligence has predictive validity "in domains such as academic performance, job performance, negotiation, leadership, emotional labor, trust, work-family conflict, and stress."[37] While some contend that emotional intelligence and personality are the same, other studies reveal that emotional intelligence is measuring something apart from personality.[38] Specifically, when measuring emotional intelligence as a separate construct, it can be measured separately from intelligence and personality."[39] In one 1995 study, it was claimed that emotional intelligence was the most significant job performance predictor.[40] However, as in many areas of research, the keynote finding of one study does not even make the footnote of a similar study. Such was the case in 2011 when a study, relying on much more data than the 1995 sample, could not support the earlier claim that EQ predicts job performance.[41] Although the exact role EQ plays in the workplace is still up for debate, it is reasonable to assume from the multitude of studies linking EQ to various performance factors that a valid and reliable emotional intelligence test used in selection process should result in useful data.

Applicant Faking

To the extent that personality and EQ tests are used in hiring, the issue of applicant faking needs to be addressed. Faking is defined "as the tendency to deliberately present oneself in a more positive manner than is accurate in order to meet the perceived demands of the testing situation."[42] The concern is that a person with high cognitive abilities will have the intellectual skill necessary to identify the answers that will maximize their chances of getting a position. A quick search on the Internet will find advice on how to fake these tests. One article, geared toward lawyers seeking employment with firms who conduct personality or EQ tests, notes:

I'm not convinced that you can't 'game' the test to some extent. So here are my tips for 'passing' the test:

- Resist the urge to be too revealing. The assessment is part of the job interview, not something for your own enlightenment. If you are curious about your psychological profile, take one of the tests out there on your own dime.

- Be a social animal. If you need to lock yourself in a soundproof room to do your work, don't admit it. These days, law firms are very keen on team work. Never mind that most of the big rainmakers tend to be solipsistic egomaniacs. The buzz word is 'cooperation."

- Be sunny. Lawyers are paid to look at the worst-case scenarios, so they tend to be skeptical, if not pessimistic. Despite your inclination to look on the dark side, try to project a positive, 'I'll-find-a-solution' attitude. That's what clients want to hear.

- Be cool. If you get angry or take criticism badly, don't admit it. Grit your teeth and say you welcome criticism—and that you always learn from it.

- Review math. Yes, there was a math section on the test that completely threw me. It might help to buy one of those SAT prep books.[43]

One recent study found faked answers for one quarter to one half of the applicants.[44] So how can employers who want to use personality or EQ tests in their selection process mitigate against the risk of applicant faking? Counter-measures to faking include the test and retest approach to see if an individual is consistent in their answers, or asking questions that require quick responses.[45] But counter-measures to faking may result in less reliable and valid results since some tools used to detect faking do not work well.[46]

Skepticism in Personality Testing

There are some skeptics in the general population who are derisive of these tests because they feel the questions posed in them are irrelevant to determining a person's personality or emotional intelligence. For example, one exam used in selecting first year legal associates asks "do you like flowers?"[47] Clearly an applicant's affection for flowers is not connected to the knowledge, skills, or abilities necessary to be a successful lawyer. It is this type of question that skeptics use to prove, at least to themselves, the total irrelevancy of psychological testing. However, proponents of these tests would say these cynics are wrong because they misunderstand the purpose behind the question. Personality tests may ask a series of irrelevant questions because the test is examining the patterns behind the responses, not the answer to any particular question—it is that pattern that provides insight into the test taker's personality.

Legal Considerations

As more and more companies decide to utilize personality and emotional intelligence tests in the employee selection process, applicant faking and placating skeptics are not the only hazards a company can expect. If not constructed properly, the potential legal ramifications of these tests can be massive. The two most significant legal considerations in using personality and emotional intelligence tests are Title VII discrimination and

discrimination under the Americans with Disabilities Act ("ADA"). While intentional discrimination is certainly possible, the more likely risk for companies acting in good faith involves inadvertent discrimination through the use of valid and reliable instruments.

Title VII Discrimination and Validation Studies

The Federal Civil Rights Act of 1964 generally prohibits employers from discriminating on the basis of "race, color, religion, sex, or national origin" in the employment context, including the employee selection process.[48] To assist employers in the selection process, Title VII allows professionally developed ability tests as long as they are not "designed, intended or used to discriminate because of race, color, religion, sex or national origin."[49] Personality differences between races are small and should not impact the use of personality tests in the employee selection process.[50] In the first Supreme Court case that examined unintentional discrimination, *Griggs v. Duke Power Co.*, the Court accepted a lower court finding that that the business was not intentionally discriminating against the plaintiffs based on race. The Court then shifted its inquiry to the employer's use of two commercially available ability tests[51]—both still in use today—and held that these facially non-discriminatory tests violated Title VII because the tests had a disparate impact on the African-American plaintiffs and the employer did not prove that the tests were related to job performance.[52] The *Griggs* Court, however, ended its opinion with agreement that employee selection tools are extremely important to business, but that business needs to use tests that are designed "for the job and not the person in the abstract."[53] Presumably, if the employer in *Griggs* had conducted a meaningful study and determined that the two ability tests were related to job performance, then the Court would have found there was no Title VII violation.[54] Today, the Court's jurisprudence has been codified into Title VII. To prevail in a disparate impact case, a plaintiff must establish that at least one of two tests has been violated. The first test requires the plaintiff to prove that an employment practice results in disparate impact which, if proven, shifts the burden to the defendant to demonstrate that the practice in question is consistent with business necessity.[55] The second test requires the plaintiff to prove that there was an alternative employment practice, the defendant refused to adopt it, and the alternative employment practice would have eliminated or reduced the disparate impact.[56] Presumably, the employer must also have been aware of the alternate employment practice at the time the defendant was being considered for employment.[57] Though most of the litigation involving alternative employment practices involves the use of employment tests, plaintiffs have rarely prevailed because their suggested alternatives were neither less discriminatory nor advanced the employer's purpose in using the test.[58] This leaves the first test—job relatedness—as the only significant disparate impact issue facing legal employers that use personality tests.

A disparate impact claim is, basically, a plaintiff proving discrimination through the use of statistics. An employer can then defeat a disparate impact claim by "proving business necessity, bona fide occupational qualifications, or validity."[59] The bona fide occupational qualification defense only applies to sex and religious discrimination and therefore only applies to a small group of employers.[60] Business necessity is limited to

safety concerns for those in the protected class (e.g., prohibiting pregnant women from working on a job that would exposes them to lead, which would be dangerous for the unborn child).[61] This leaves employers with the need to establish validity for their selection tools. To help government agencies and employers with a uniform understanding of validation, in 1978 the government created the Uniform Guidelines on Employee Selection Procedures ("Guidelines").[62] The Guidelines provide options for establishing validity, though modern science is often opposed to the older science enshrined in the Guidelines.[63] In one recent case rejecting disparate impact, the Supreme Court held that the City of New Haven, Connecticut had developed an examination that was job related, was necessary for the firefighting business at issue in the case, and had the requisite validity.[64] This demonstrates the importance of validating tests before administering them.

Americans with Disabilities Act

The Americans with Disabilities Act (ADA) prohibits employers from conducting pre-employment medical exams.[65] Though most employers are only interested in identifying personality traits necessary for a particular position, some personality tests might also have the ability to identity a medical condition, thereby violating the ADA. For example, in *Karraker v. Rent-A-Center*, the Court held that a personality test that could have been used by the employer to diagnose a medical condition violated the ADA. Specifically, the employer used the Minnesota Multiphasic Personality Inventory ("MMPI), which can measure "depression, hypochondriasis, hysteria, paranoia, and mania."[66] The Court rejected the "we aren't using it for that" argument and explained that because the test can reveal mental illness then it should be legally classified as a medical exam."[67] In another case, an employer asked candidates whether they agreed or disagreed with the following statements:

- People do a lot of things that make you angry.
- There's no use having close friends; they always let you down.
- Many people cannot be trusted.
- You are unsure of what to say when you meet someone.[68]

The applicants were concerned that the questions might identify mental illness, which is prohibited by the ADA, so the company agreed to remove the questions from future tests.[69] Personality tools that are designed by knowledgeable psychologists familiar with employment laws should have no difficulty in avoiding an ADA violation.

Conclusion

Making poor hiring decisions not only has the potential to create a toxic workplace environment, but it can be expensive. Each bad hire costs a business 1.5 times[70] to 5 times that employee's salary and benefits.[71] Assuming a $50,000 combined salary and benefits, the bad hire will cost an employer at least an additional $75,000. Even though an employer may be challenged in court for using personality and EQ tests in employee selection, the benefits of more successful employees far outweigh potential legal costs. The key is for employers to use valid, reliable, and legally sustainable tests in hiring

employees, not only because this will reduce potential lawsuits but also because it is the only way that employers can scientifically identify the best candidates for the job. ℵ

H. Beau Beaz currently serves as the Associate Dean for Academic Effectiveness & Professor of Law at the Charlotte School of Law. Previously, Baez was the director of the Tax Law program at Concord University School of Law and counsel for the Multistate Tax Commission. He received both a J.D. and a Master of Laws in Taxation from Georgetown University Law Center and was a law clerk for the United States Attorney's Office.

---

[1] Ruth Mantell, *Job Seekers, Get Ready for Personality Tests: More Employers are using Pre-Hire Assessments*, MARKET WATCH (September 12, 2011) http://www.marketwatch.com/story/job-seekers-get-ready-for-personality-tests-2011-09-12.
[2] *Id.* at 491.
[3] WAYNE WEITEN, PSYCHOLOGY: THEMES AND VARIATIONS 385 (8th ed. 2010).
[4] Frank L. Schmidt & John E. Hunter, *The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings*, 124 PSYCHOLOGICAL BULLETIN 262 (1998).
[5] Lee Borghans, Angela Duckworth, et al., *The Economics and Psychology of Personality Traits* 42 Nat'l Bureau of Econ. Research, Working Paper No. 13810 (2008).
[6] ANNIE MURPHY PAUL, THE CULT OF PERSONALITY: HOW PERSONALITY TESTS ARE LEADING US TO MISEDUCATE OUR CHILDREN, MISMANAGE OUR COMPANIES, AND MISUNDERSTAND OURSELVES (2004).
[7] *Id.*
[8] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 135 (14th ed. 2011).
[9] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[10] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[11] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[12] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[13] STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[14] R. R. MCCRAE & P. T. COSTA, PERSONALITY IN ADULTHOOD: A FIVE-FACTOR THEORY PERSPECTIVE (2003); STEPHEN P. ROBBINS AND TIMOTHY A. JUDGE, ORGANIZATIONAL BEHAVIOR, 138 (14th ed. 2011).
[15] ROBERT D. GATEWOOD, HUBERTO S. FIELD, AND MURRAY BARRICK, HUMAN RESOURCE SELECTION, 507 (7th ed., 2011).
[16] Self-report questionnaires are less expensive to administer, which is why they tend to be more popular than approaches that require the administration by a trained psychologist.
[17] ROBERT D. GATEWOOD, HUBERTO S. FIELD, AND MURRAY BARRICK, HUMAN RESOURCE SELECTION, 511 (7th ed., 2011).
[18] Hogan Assessments, Hogan Personality Inventor: Overview Guide, http://www.hoganassessments.com/sites/default/files/assessments/pdf/HPI_Brochure.pdf (The HPI utilizes the following seven dimensions:
"Adjustment: confidence, self-esteem, and composure under pressure
Ambition: initiative, competitiveness, and desire for leadership roles
Sociability: extraversion, gregarious, and need for social interaction
Interpersonal Sensitivity: tact, perceptiveness, and ability to maintain relationships
Prudence: self-discipline, responsibility and conscientiousness
Inquisitive: imagination, curiosity, and creative potential
Learning Approach: achievement-oriented, stays up-to-date on business and technical matters"
[19] Joyce Hogan and Brent Holland, Using Theory to Evaluate Personality and Job-Performance Relations: A Socioanalytic Perspective, 88 JOURNAL OF APPLIED PSYCHOLOGY 100, 103 (2003).

[20] *See generally*, Frederick P. Morgeson, et al., *Are We Getting Fooled Again? Coming to Terms with Limitations in the Use of Personality Tests for Personnel Selection*, 60 PERSONNEL PSYCHOLOGY 1029 (2007).
[21] Wesley A. Scroggins, Steven L. Thomas, and Jerry A. Morris, *Psychological Testing in Personnel Selection, Part I: A Century of Psychological Testing*, 38 PUBLIC PERSONNEL MANAGEMENT 99, 105 (2008).
[22] E. E. Ghiselli and R. P. Barthol, *The Validity of Personality Inventories in the Selection of Employees*, 38 JOURNAL OF APPLIED PSYCHOLOGY 18 (1953).
[23] *See generally*, E. E. Ghiselli and R. P. Barthol, *The Validity of Personality Inventories in the Selection of Employees*, 38 JOURNAL OF APPLIED PSYCHOLOGY 18, 20 (1953).
[24] *See generally*, Robert M. Guion and Richard F. Gottier, Validity of Personality Measures in Personnel Selection, 18 Personnel Psychology 135, 141 (1965).
[25] Robert M. Guion and Richard F. Gottier, *Validity of Personality Measures in Personnel Selection*, 18 PERSONNEL PSYCHOLOGY 135, 159-160 (1965).
[26] Frederick P. Morgeson, et al., *Are We Getting Fooled Again? Coming to Terms with Limitations in the Use of Personality Tests for Personnel Selection*, 60 PERSONNEL PSYCHOLOGY 1029, 1033 (2007).
[27] Frederick P. Morgeson, et al., *Are We Getting Fooled Again? Coming to Terms with Limitations in the Use of Personality Tests for Personnel Selection*, 60 PERSONNEL PSYCHOLOGY 1029, 1037 (2007).
[28] Huy Le, In-Sue Oh, Steven B. Robbins, et al., *Too Much of a Good Thing: Curvilinear Relationships Between Personality Traits and Job Performance*, 96 JOURNAL OF APPLIED PSYCHOLOGY 113 (2011).
[29] Huy Le, In-Sue Oh, Steven B. Robbins, et al., *Too Much of a Good Thing: Curvilinear Relationships Between Personality Traits and Job Performance*, 96 JOURNAL OF APPLIED PSYCHOLOGY 113, 129 (2011).
[30] WAYNE WEITEN, PSYCHOLOGY THEMES AND VARIATIONS 361-362 (8th ed. 2010) ("An intelligence quotient (IQ) is a child's mental age divided by chronological age, multiplied by 100.").
[31] *Id.* at 364, 366.
[32] Peter Salovey & John Mayer, *Emotional Intelligence*, 9 IMAGINATION, COGNITION, AND PERSONALITY 185 (1990).
[33] WAYNE WEITEN at 386.
[34] Seventy-One Percent of Employers Say They Value Emotional Intelligence Over IQ, According to CareerBuilder Survey, August 18, 2011, http://m.prnewswire.com/news-releases/seventy-one-percent-of-employers-say-they-value-emotional-intelligence-over-iq-according-to-careerbuilder-survey-127995518.html
[35] Id.
[36] Id.
[37] Ernest H. O'Boyle et al., *The Relation Between Emotional Intelligence and Job Performance: A Meta-Analysis*, 32 J. ORGANIZ. BEHAV. 789 (2010) (citations omitted).
[38] J. C. Rode et al., *Emotional Intelligence and Individual Performance: Evidence of Direct and moderated Effects*, 28 JOURNAL OF ORGANIZATIONAL BEHAVIOR 399 (2007).
[39] O'Boyle at 806.
[40] D. GOLEMAN, EMOTIONAL INTELLIGENCE: WHY IT CAN MATTER MORE THAN IQ (1995).
[41] O'Boyle at 804.
[42] Jinyan Fan, Dingguo Gao, Sarah A. Carroll, et. al., *Testing the Efficacy of a New Procedure for Reducing Faking on Personality Tests Within Selection Contexts*, AMERICAN JOURNAL OF APPLIED PSYCHOLOGY 1, 2 (2012).
[43] Vivia Chen, *The Careerist Goes on the Couch*, THE CAREERIST (Feb. 1, 2011). http://thecareerist.typepad.com/thecareerist/2011/02/aceing-the-psych-test.html.
[44] R. L. Griffith, et al, *Do Applicants Fake? An Examination of the Frequency of Applicant Faking Behavior*, 36 PERSONNEL REVIEW, 341 (2007).
[45] Jinyan Fan, Dingguo Gao, Sarah A. Carroll, et. al., *Testing the Efficacy of a New Procedure for Reducing Faking on Personality Tests Within Selection Contexts*, AMERICAN JOURNAL OF APPLIED PSYCHOLOGY 1, 2 (2012).
[46] Mitchell H. Peterson, Richard L. Griffith, Joshua A. Isaacson, et. al., *Applicant Faking, Social Desirability, and the Prediction of Counterproductive Work Behaviors*, 24 HUMAN PERFORMANCE 270, 286 (2011).

[47] Vivia Chen, *The Careerist Goes on the Couch*, THE CAREERIST (Feb. 1, 2011) http://thecareerist.typepad.com/thecareerist/2011/02/aceing-the-psych-test.html
[48] 42 U.S.C.A. § 2000e-2(a)(1). The shorthand term "Title VII," used by practitioners and in the literature, refers to this section's location in Pub.L. 88-352 passed in 1964 rather than to its location in Title 42, Subchapter VI of the U.S. Code.
[49] 42 U.S.C.A. §2000e-2(h).
[50] Hannah J. Foldes, Emily E. Duehr, and Deniz S. Ones, Group Differences in Personality: Meta-Analyses Comparing Five U.S. Racial Groups, 61 Personnel Psychology, 579, 605 (2008).
[51] The Wonderlic Personnel Test, today called the Wonderlic Classic Cognitive Ability Test, claims to "measure a candidate's ability to understand instructions, learn, adapt, solve problems and handle the mental demands of the position." http://www.wonderlic.com/assessments/ability/cognitive-ability-tests/classic-cognitive-ability-test (last visited July 4, 2012). The Bennett Mechanical Comprehension Test is used for assessing mechanical aptitude, "with a focus on spatial perception and tool knowledge rather than manual dexterity." http://www.pearsonassessments.com/HAIWEB/Cultures/en-us/Productdetail.htm?Pid=015-8341-430&Mode=summary (last visited July 4, 2012).
[52] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).
[53] Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).
[54] *See*, Griggs v. Duke Power Col, 401 U.S. 424, 431 (1971).
[55] 42 U.S.C.A. § 2000e-2(k)(1)(A)(i).
[56] 42 U.S.C.A. § 2000e-2(k)(1)(A)(ii).
[57] 2 Lex K. Larson, EMPLOYMENT DISCRIMINATION §24.01, at 24-5 (2d ed. 2012).
[58] 2 Lex K. Larson, EMPLOYMENT DISCRIMINATION §24.01, at 24-8 (2d ed. 2012).
[59] ROBERT D. GATEWOOD, HUBERTO S. FIELD, AND MURRAY BARRICK, HUMAN RESOURCE SELECTION, 39 (7th ed., 2011).
[60] Id.
[61] Id.
[62] Created by the Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and the Department of Justice.
[63] Robert M. Guion and Scott Highhouse, Essentials of Personnel Assessment and Selection, 87 (2006).
[64] Ricci v. DeStefano, 557 U.S. 557, 587-588 (2009).
[65] 42 U.S.C.A. § 12112(d)(2)(A).
[66] Karraker v. Rent-A-Center, Inc., 411 F.3d 831, 833 (7th Circuit, 2005).
[67] *Id.* at 837.
[68] Eve Tahmincioglu, *Employers Turn to Tests to Weed Out Job Seekers: Some Screens May Violate Law, But Applicants Rarely Have Choice*, Today Money, 8/15/2011 http://today.msnbc.msn.com/id/44120975/ns/today-money/t/employers-turn-tests-weed-out-job-seekers/
[69] *Id.*
[70] *See* Kay Lazar, *Employers Test with a New Attitude: Controversial Questionnaires Screen Applicants for Hire Purposes*, Boston Herald, Apr. 18, 1999, at 3.
[71] Survey: Bad Hires Cost Big Money, Philadelphia Business Journal (April 11, 2006) http://www.bizjournals.com/philadelphia/stories/2006/04/10/daily19.html