Prior to beginning work on this assignment, please read the required textbook chapters and articles for this week.

Create a PowerPoint presentation with 16 to 20 slides (not including the title and reference slides) entitled *Ethical and Professional Issues in Psychological Testing*. Your presentation must provide 2 to 3 slides for each of the required topics and include appropriate citations of your referenced sources. Separate reference slides, which follow APA formatting guidelines for a References page, must be included at the end of the presentation. You must create your own template and organize your presentation in the sequence provided. Do not use a font smaller than 20 pt. You are encouraged to insert relevant figures and graphics. Make sure to appropriately cite any images you use. If you include a table or figure from a journal article, cite it according to APA guidelines. The notes section of each slide must include the text for oral comments you would make while presentating the materials to a live audience.

References must be cited according to APA guidelines as outlined in the Ashford Writing Center (Links to an external site.). For assistance with creating a visually engaging and readable presentation, you may review Garr Reynolds's tips for creating presentations (Links to an external site.).

The presentation must cover each of the following topics in the order presented below.

The Ethical and Social Implications of Testing

- Provide an overview and brief evaluation of the ethical and social implications of psychological assessment.

Professional Responsibilities

- Describe the responsibilities of both test publishers and test users.

Testing Individuals Representing Cultural and Linguistic Diversity

- Analyze and describe issues related to the testing of cultural and linguistic minorities.

Reliability

- Explain the common sources of measurement error and how measurement error can impact reliability.

Validity

- Create a diagram or figure to compare the types of validity discussed in the textbook.
- Describe the extravalidity concerns related to testing.

- Review the articles by Fergus (2013), Kosson, et al. (2013) and Mathieu, Hare, Jones, Babiak, & Neumann (2013). Analyze the information presented in these articles on factor analysis and describe how it is used to validate the constructs of the instruments.

Clinical Versus Statistical Prediction

- Compare clinical and statistical prediction of mental health decisions based on the work of Ægisdóttir, et al. (2006) and Grove & Lloyd (2006).

Application One: An Ethical and Professional Quandry

- Select one of the Ethical and Professional Quandries in Testing from Case Exhibit 1.2 in your textbook and describe the ethical issues specific to the scenario you selected. Include an analysis of the relevant principles from Standard 9 in the APA Ethical Principles of Psychologists and Code of Conduct (Links to an external site.)
- Taking on the role of the psychologist or counselor in the chosen scenario, describe how you might respond to the challenge you selected and provide a brief rationale for your decision.

Application Two: Evidence-Based Medicine

- Summarize Youngstrom's (2013) recommendations for linking assessment directly to clinical decision making in evidence-based medicine.
- Elaborate on each of Youngstrom's recommendations by providing practical examples that illustrate the relevance of the recommendations in a clinical setting.

Application Three: Selecting Valid Instruments

- Create a research hypothesis or brief clinical case scenario in which you must select an instrument to measure intolerance for uncertainty.
- Use the information in the Fergus (2013) article to support which measure to use.

The presentation

- Must consist of 16 to 20 slides (not including title and reference slides) that are formatted according to APA style as outlined in the Ashford Writing Center (Links to an external site.).
- Must include a separate title slide with the following:
  - Title of presentation
  - Student's name
  - Course name and number
  - Instructor's name
  - Date submitted

- Must use the assigned chapters in the course text, Standard 9 from the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct, and the 3 required peer-reviewed articles assigned for Week One.
- Must document all sources in APA style as outlined in the Ashford Writing Center.
- Must include separate reference slides formatted according to APA style as outlined in the Ashford Writing Center.

• **Major Contribution**

The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction

Stefanía Ægisdóttir
Michael J. White
Paul M. Spengler
Alan S. Maugherman
Linda A. Anderson
Robert S. Cook
Cassandra N. Nichols
Georgios K. Lampropoulos
Blain S. Walker
Genna Cohen
Jeffrey D. Rush
Ball State University

Clinical predictions made by mental health practitioners are compared with those using statistical approaches. Sixty-seven studies were identified from a comprehensive search of 56 years of research; 92 effect sizes were derived from these studies. The overall effect of clinical versus statistical prediction showed a somewhat greater accuracy for statistical methods. The most stringent sample of studies, from which 48 effect sizes were extracted, indicated a 13% increase in accuracy using statistical versus clinical methods. Several variables influenced this overall effect. Clinical and statistical prediction accuracy varied by type of prediction, the setting in which predictor data were gathered, the type of statistical formula used, and the amount of information available to the clinicians and the formulas. Recommendations are provided about when and under what conditions counseling psychologists might use statistical formulas as well as when they can rely on clinical methods. Implications for clinical judgment research and training are discussed.

A large portion of a counseling psychologist's work involves deciding what information to collect about clients and, based on that information, predicting future client outcomes. This decision making can occur both at the microlevel, such as moment-to-moment decisions in a counseling session, and at the macrolevel, such as predictions about outcomes such as suicide risk, violence, and response to treatment (Spengler, Strohmer, Dixon, & Shivy, 1995). Because the quality of client care is often determined

THE COUNSELING PSYCHOLOGIST, Vol. 34 No. 3, May 2006 341-382
DOI: 10.1177/0011000005285875

by the accuracy of these decisions (Dawes, Faust, & Meehl, 1989; Meyer et al., 1998; Spengler, 1998), determining the best means for decision making is important.

Two major approaches to decision making have been identified: the clinical and the statistical, which is also called mechanical (Dawes et al., 1989). Clinical prediction refers to any judgment using informal or intuitive processes to combine or integrate client data. Psychologists use the clinical method when their experience, interpersonal sensitivity, or theoretical perspective determines how they recall, synthesize, and interpret a client's characteristics and circumstances.

Such intuitive or "gut-level" inferences are greatly reduced in the statistical approach. Predictions are based on empirically established relations between client data and the condition to be predicted (Dawes et al., 1989). A psychologist who declares that his or her clinical impression suggests a client may be suicidal has used the clinical method. By contrast, when using the statistical method, client data are entered into formulas, tables (e.g., actuarial tables), or charts that integrate client information with base rate and other empirical information to predict suicide risk. While the statistical method is potentially 100% reproducible and well specified, the clinical method is neither as easily reproduced nor as clearly specified (Grove, Zald, Lebow, Snitz, & Nelson, 2000).

Meehl (1954) contended that while the clinical method requires specific

---

Alan S. Maugherman is now in private practice at the Center for Psychological Development, Muncie, Indiana. Linda A. Anderson is now at the University Counseling and Psychological Services, Oregon State University. Robert S. Cook is now in private practice at Lifespan, North Logan, Utah. Cassandra N. Nichols is now at the Counseling and Testing Services, Washington State University. Blain S. Walker is now at the Tripler Army Medical Center, Honolulu, Hawaii. Genna R. Freels is now at the Louisiana Professional Academy, Lafayette, Louisiana. Funding for this project was provided to Paul M. Spengler by Grant MH56461 from the National Institute of Mental Health, Rockville, Maryland; by six grants to Paul M. Spengler from the Internal Grants Program for Faculty, Ball State University, Muncie, Indiana (two summer research grants, two summer graduate research assistant grants, an academic year research grant, and a new faculty research grant); and by three Lyell Bussell summer graduate research assistant grants, Ball State University, Muncie, Indiana. Preliminary findings from the clinical versus statistical prediction meta-analysis were presented at the annual meetings of the American Psychological Association in Washington, D.C., August 2000; Boston, August 1999; San Francisco, August 1998; Toronto, Ontario, Canada, August 1996; and New York City, August 1995, and by invited address at the annual meeting of the Society for Personality Assessment, New Orleans, March 1999. The authors extend special thanks to Kavita Ajmere, Corby Bubp, Jennifer Cleveland, Michelle Dorsey, Julie Eiser, Layla Hunton, Christine Look, Karsten Look, K. Christopher Rachal, Teresa Story, Marcus Whited, and Donald Winsted III for instrumental assistance in obtaining articles, coding studies, and managing data for the project. Correspondence concerning this article should be addressed to Stefanía Ægisdóttir, Department of Counseling Psychology, Teachers College 622, Ball State University, Muncie, IN 47306; e-mail: stefaegis@bsu.edu.

training, the statistical method does not. The statistical method requires only inserting data into a formula specifically designed for a particular judgment task. This may not be entirely true. Despite the use of formulas or tables to integrate information, the statistical method may require advanced training in the collection of relevant clinical and research-based information. Furthermore, advanced training may enhance a clinician's perceptions, which in turn may be quantified and used in a statistical model. For example, a clinician may believe a client has the potential for suicide, translate this impression into a number on a rating scale, and then statistically combine this number with other data to predict the client's risk for suicide (e.g., Westen & Weinberger, 2004).

To determine how counseling psychologists can be most effective in their decision making, knowing when and under what conditions each method is superior is important. The purpose of our meta-analysis is to articulate this knowledge.

THE CLINICAL VERSUS STATISTICAL PREDICTION CONTROVERSY

The search for the most accurate decision-making method is not new. In fact, this question has been debated for more than 60 years (Dawes et al., 1989; Meehl, 1954). The debate began with Meehl's (1954) book *Clinical Versus Statistical Prediction,* in which Meehl theoretically analyzed the relation between the clinical and statistical methods of prediction and summarized findings from existing literature. Meehl found that in all but 1 of 20 studies, statistical methods were more accurate than or equally accurate as the clinical method. He concluded that clinicians' time should be spent doing research and therapy, whereas work involving prognostic and classification judgments should be left to statistical methods.

Holt (1958), the most adamant defender of the clinical method, criticized Meehl's (1954) conclusions. Holt's critique involved essentially two issues: (a) the identification and assessment of predictive variables and (b) how they should be integrated. Holt believed that Meehl had given insufficient attention to the sophistication with which clinicians identify the criteria they are predicting, what variables to use in their prediction, and the strength of the relationship between predictors and criteria. In Holt's view, clinicians can identify these variables only through training and experience with comparable cases. After identifying the relevant variables, they are assessed. Assessment may be as much qualitative as quantitative. Holt's second criticism was that Meehl pitted "naïve clinical integration" of prediction against statistical decision making. A fairer comparison would compare statistical methods with "sophisticated clinical decision making and integration" (Holt,

1958) According to Holt, sophisticated clinical decision making is based on sophisticated data. These data are both qualitative and quantitative, have been gathered in a systematic manner, and have known relationships with what is being predicted. Unlike the statistical approach, the clinician remains the prime instrument, combining the data and making predictions that are tailored to each person. Holt presented data suggesting a superiority for sophisticated clinical rather than statistical procedures in predicting success in clinical training. On the basis of these findings, Holt argued for a combination of clinical and statistical methods (i.e., sophisticated clinical) that would be systematic and controlled and sensitive to individual cases.

Since this time, other narrative and box-score reviews of the literature on the differential accuracy of clinical and statistical methods have been published (e.g., Dawes et al., 1989; Garb, 1994; Grove & Meehl, 1996; Kleinmuntz, 1990; Russell, 1995; Sawyer, 1966; Wiggins, 1981). Narrative reviews are traditional literature reviews; box-score reviews count statistical significance and summarize studies in a table format. These reviews nearly always supported Meehl's (1954) conclusion that statistical methods were more accurate than or, at minimum, equally as accurate as clinical prediction methods (for a rare exception, see Russell, 1995). A recent meta-analysis of the clinical versus statistical literature (Grove et al., 2000) also supported earlier findings. Grove et al. (2000) found a consistent advantage ($d = .12$) for statistical prediction over clinical prediction across various types of nonmental health and mental health predictors and criteria.

Influence of the Statistical Versus Clinical Prediction Controversy

Despite the repeated conclusion that statistical prediction methods are more accurate than clinical procedures, the findings have had little influence on clinical practice (Dawes et al., 1989; Meehl, 1986). Dawes et al. (1989) and Meehl (1986) offered several reasons for this. They suggested that clinicians lack familiarity with the literature on clinical versus statistical prediction, are incredulous about the evidence, or believe that the comparisons were procedurally biased in favor of statistical prediction methods. They also proposed that certain aspects of education, training, theoretical orientation, and values might influence their reluctance to recognize advantages associated with statistical decision methods. Most clinicians highly value interpersonal sensitivity. Because of this, some may believe that the use of predictive formulas dehumanizes their clients. A corollary is that the use of group-based statistics or nomothetic rules is inappropriate for any particular individual. Practitioners are also subject to confirmatory biases such that they recall instances in which their predictions were correct but fail

to recall those instances in which statistical prediction was more accurate. One might add another reason: Some accounts have simply been too broad to convince mental health practitioners. In some instances (e.g., Grove et al., 2000), the literature that has reviewed clinical versus statistical prediction includes research and criteria that range from mental health to medicine to finance.

Use of Statistical Prediction Models

Perhaps as a result of the limited influence of clinical versus statistical comparison studies, few statistical prediction models are available to counseling psychologists and psychotherapy practitioners (Meyer et al., 1998). Clinicians working in forensic settings, however, have developed such models. In fact, numerous funded research projects have been conducted to aid in classifying juvenile and adult prison inmates (e.g., Gottfredson & Snyder, 2005; Quinsey, Harris, Rice, & Cormier, 1998; Steadman et al., 2000; Sullivan, Cirincione, Nelson, & Wallis, 2001). One such effort is the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 1998), which is a statistical system for predicting recidivism of imprisoned violent offenders.

The VRAG is based on more than 600 Canadian maximum security inmates who were released either back to the community, to a minimum security hospital, or to a halfway house. After a series of correlation analyses of predictor and outcome variables, a set of stepwise regression models was conducted. These analyses reduced the original 50 predictors to 12. These include psychopathy checklist scores (Hare, 1991), elementary school maladjustment scores, presence of a personality disorder, age at time of offense, separation from parents at an age younger than 16, failure on prior conditional release, nonviolent offense history score (using an instrument), marital status, schizophrenia diagnosis, most serious injury of offender's victim, alcohol abuse score, and gender of offender's victim. Each predictor was assigned a specified weight based on the empirical relationship with the outcome variable. Summing the resultant scores yields a probability estimate for an offender's future violence within the next 7 and 10 years. For instance, scores between +21 and +27 indicate a 76% likelihood for future violence, whereas scores between –21 and –15 suggest a probability of only 8%. The authors have validated this model for different groups of inmates (e.g., arsonists or sex offenders), with promising results (see Quinsey et al., 1998, for more detailed use of this statistical model).

In addition to forensics, statistical prediction formulas have been developed to aid with student selection for undergraduate, graduate, and professional schools. As an example, Swets et al. (2000) described a statistical prediction formula used in selecting candidates at the University of Virginia

School of Law. This formula consists of four predictor variables: undergraduate grade point average (GPA), mean GPA achieved by students from the applicants' college, scores from the Law School Admissions Test (LSAT), and the mean LSAT score achieved by all students from the applicants' college. Scores from these predictors are combined into a decision index of which a specific score indicates a threshold for admission. This statistical prediction formula predicts grades for 1st-year students and is used in combination with variables that are harder to quantify to select students (cf. Swets et al., 2000). Harvey-Cook and Taffler (2000) developed a statistical model using biographical data, frequently found on application forms and resumes, to predict success in accounting training in the United Kingdom. This six-variable model was developed on 419 accounting trainees. Retesting it on an independent sample of 243 trainees, Harvey-Cook and Taffler showed that their model could classify 88% of those failing and 33% of those successful in accounting training. The authors concluded that their model delivered better and more cost-effective results than clinical judgment methods currently used for this purpose in the United Kingdom (Harvey-Cook & Taffler, 2000).

Test cutoff scores offer another instance of a statistical procedure that may aid clinical decision making. Indeed, cutoff scores may be more readily available and easily constructed than statistical formulas. As an example, three Minnesota Multiphasic Personality Inventory–2 (MMPI-2) scales have been useful in classifying substance abuse: MacAndrew Alcoholism–Revised (MAC-R), Addiction Potential Scale (APS), and Addiction Acknowledgment Scale (AAS) (Rouse, Butcher, & Miller, 1999; Stein, Graham, Ben-Porath, & McNulty, 1999). Relying on data from 500 women and 333 men seeking outpatient mental health services, Stein et al. (1999) found that cutoff scores on the MAC-R correctly classified 86% of the women and 82% of the men as either substance abusers or nonabusers. In the case of the AAS, cutoff scores could predict 92% of women and 81% of men as either substance abusers or nonabusers. Likewise, cutoff scores with the APS enabled accurate prediction of 84% of women and 79% of men as either abusing or not abusing substances. This method of classification greatly exceeds the base rates for chance classification. For women, the positive predictive power (ability to detect substance abusers) for MAC-R, AAS, and APS was 100%, 79%, and 53%, respectively. These values compare with a base rate of 16%. For men, the respective positive predictive power for MAC-R, AAS, and APS was 100%, 68%, and 77%, respectively, which compare with a base rate of 27%.

Purpose of This Meta-Analysis

The current meta-analysis seeks to address several omissions in the literature on clinical versus statistical prediction. Although Grove et al.'s (2000)

important study confirmed prior conclusions about the relative merits of clinical and statistical prediction methods, questions still remain regarding the application of the findings to judgment tasks commonly encountered by mental health practitioners. First, their review combined literature from psychology, medicine, forensics, and finance. Consequently, conclusive results are not provided about prediction accuracy for mental health clinical and counseling practitioners relative to statistical methods. Second, even though Grove et al. examined the influence of various study design characteristics (e.g., type of criterion, professional background of clinical judges, judge's level of experience, and amount of data available to the judges versus the statistical formulas), the influence of these design characteristics on the accuracy of prediction was not investigated when the criteria were psychologically related. Instead, Grove et al. investigated the influence of these study design variables on the overall effect, including studies from the diverse professional fields listed earlier. Similarly, despite Grove et al.'s examination of the influence of criterion type on the overall effect of the difference between clinical and statistical prediction accuracy, their criteria breakdown was broad (i.e., educational, financial, forensic, medical, clinical-personality, and other). The breakdown offers little specific information on which counseling psychologists can rely to decide when and under what conditions they should use clinical or statistical methods.

The first aim of this meta-analysis was to synthesize studies that had examined the differential accuracy of clinical and statistical judgments in which the prediction outcome was relevant to counseling psychology. Second, we examined studies in which predictions by mental health professionals were compared with statistical methods. In a typical study comparing these two methods, clinicians first synthesized client data (e.g., interview data, psychological tests, or a combination of interview information and one or more psychological tests) and then made a classification judgment (e.g., diagnosis) or predicted some future outcome (e.g., prognosis). The accuracy of these judgments was compared with a statistical prediction scheme in which the same (sometimes less or more) information was entered into a statistical formula that had been previously designed on the basis of empirical relations between the predictors (specific client data) and the criterion (the prediction task of interest). Third, we examined questions generated from the years of debate about the relative merits of clinical and statistical prediction. More specifically, we examined how the differential accuracy between clinical and statistical methods was affected by (a) type of prediction, (b) setting from which the data were gathered, (c) type of statistical formula, (d) amount of information provided to the clinician and formula, (e) information provided to the clinician about base rates, (f) clinician access to the statistical formula, (g) clinician expertness, (h) our evaluation

of the validity of the criteria for accurate judgment, (i) publication source, (j) number of clinicians performing predictions in a study, (k) number of criterion behaviors predicted in a study, and (l) publication year.

Meta-analyses provide detailed and comprehensive syntheses of the professional literature. As such, they are especially relevant for bridging the gap between the science of counseling psychology and how it is practiced by counseling psychologists (e.g., Chawalisz, 2003; Stricker, 2003; Wampold, 2003). The current meta-analysis addresses how counseling psychologists should best make decisions: when they should use clinical methods, when they would do well to use statistical methods, and when either is acceptable. In addition to relying on empirically supported treatment strategies, the counseling psychologist scientist-practitioner may be informed by the current meta-analysis about situations when statistical decision methods lead to more accurate clinical predictions than the clinical method.

Spengler et al. (1995), for instance, proposed an elaborated model of the scientist-practitioner, basing their clinical judgment model on Pepinsky and Pepinsky (1954). In this model, strategies were proposed to increase judgment accuracy relying on scientific reasoning. They suggested that to improve judgment accuracy, counseling psychologists (a) should be aware of their values, preferences, and expectations; (b) should use multiple methods of hypothesis testing (both confirming and disconfirming); and (c) should use judgment debiasing techniques (cf. Spengler et al., 1995). We argue that the current meta-analysis will further inform counseling psychologists as scientists not by providing information about the absolute accuracy of clinical judgment (i.e., when it may be most vulnerable to error) but instead by assessing the relative accuracy of clinical versus statistical prediction. Under conditions in which statistical prediction is superior, a successful debiasing method would use prediction methods based on empirical relations between variables (i.e., statistical methods). On the basis of this meta-analysis, we hope to also suggest options for future research and training relevant to decisions typically made by counseling psychologists.

METHOD

Study Selection

This study is part of a large-scale meta-analysis of the clinical judgment (MACJ) literature (Spengler et al., 2005). By using 207 search terms, the MACJ project identified 1,135 published and unpublished studies between

1970 and 1996 that met our criteria for inclusion in meta-analyses of mental health clinical judgment.[1] However, because of the extensive historical debate about the relative benefits of statistical versus clinical prediction, we extended our search strategy for the present study back to 1940, thus defining the current study's search period from 1940 to 1996. After an iterative process, we identified 156 studies that investigated some form of statistical prediction or model of clinical prediction for a mental health criterion compared with the accuracy of clinical judgment.

To be included in the meta-analysis, studies had to meet the following criteria: (a) a direct comparison was reported between predictions made by mental health practitioners (i.e., professionals or graduate students) and some statistical formula, (b) a psychological or a mental health prediction was made (e.g., diagnosis, prognosis, or psychological adjustment), (c) the clinicians and the statistical formula had access to the same predictor variables or cues (even though the amount of information might vary), (d) the clinicians and the formula had to make the same predictions, and (e) the studies had to contain data sufficient to calculate effect sizes. By using these selection criteria, 67 studies qualified for inclusion, yielding 92 effect sizes. When Goldberg (1965) and Oskamp (1962) were included, 69 studies produced 173 effect sizes (see below).

Specialized Coding Procedures

The MACJ project used a coding form with 122 categories or characteristics (see Spengler et al., 2005) that were grouped under the following conceptual categories: judgment task, judgment outcomes, stimulus material, clinician individual differences, standard for accuracy, method of study, and type of design. An additional coding form was constructed including study design characteristics identified in historical literature and more contemporary research as potentially affecting the differential accuracy of clinical and statistical prediction. These design characteristics became the independent variables. We also noted whether the statistical formulas were cross-validated. In this instance, cross-validated formulas refer to any statistical formulas that have been independently validated on a different sample from which the formula was originally derived. For example, if a score of 10 on an instrument developed to diagnose major depressive disorder correctly identifies 95% of persons with that disorder, to be considered a cross-validated formula (i.e., a score of 10 indicates major depression), that same score (10) had to be able to identify major depressive disorder with comparable accuracy using another sample of persons with the disorder. Coding disagreements were resolved by discussion among coders until agreement was reached.

Dependent Measure: Judgment Accuracy

The dependent variable for all analyses was judgment accuracy. For a study to be included, a criterion had to be established as the accurate judgment (e.g., prior diagnosis or arrest records). For instance, Goldberg (1970) compared clinical and statistical judgments of psychotic versus neurotic MMPI profiles to actual psychiatric diagnosis. MMPI profiles from psychiatric patients diagnosed as clearly psychotic or neurotic were presented to clinical psychologists. Their judgment about whether the MMPI profiles belonged to either a psychotic or a neurotic patient was compared with a statistical formula constructed to categorize patients as psychotic if five MMPI scales (the lie, 6 [Pa], 8 [Sc], 3 [Hy], 7 [Pt]) were elevated. These two types of judgments were compared with the prior diagnoses, which were considered the accurate judgment. In another example, Gardner, Lidz, Mulvay, and Shaw (1996) examined clinical and statistical prediction of future violence. Gardner et al. developed three statistical formulas to predict future violence on the basis of clinical (e.g., diagnosis and drug use) and demographic information as well as information about prior violence. Violence prediction based on these three models was compared with predictions made by clinicians who had access to the same information as the formulas. The accuracy of these judgments was then compared with records of violent behavior (psychiatric, arrest, or commitment records) or from patients' reports about their violent behavior. In this study, available records and patient self-reports about violent behavior served as the criteria for accurate judgment. Thus, specific criteria for accurate judgments had to be reported for a study to be included in this meta-analysis.

Effect Size Measure

As Cohen (1988) noted in his widely read book, effect sizes may be likened to the size of real differences between two groups. Estimates of effect size are thus estimates of population differences—they estimate what is really happening and are not distorted by sample size. The purpose of a meta-analysis is to estimate the effect size in a population of studies. In our case, a mean weighted effect size (d^+) was used to represent the difference between clinical and statistical prediction accuracy.[2] Effect size measured by d^+ represents the mean difference between two samples of studies expressed in standard deviation units (g) and corrected for sample size (Johnson, 1993). More specifically, the mean judgment accuracy of statistical prediction was subtracted from the mean judgment accuracy of clinical

prediction divided by the pooled standard deviation and then corrected for sample size.

In this study, the effect size (d^+) represents the magnitude, not the statistical significance, of the relative difference between clinical and statistical prediction accuracy. A negative d^+ value indicates superiority of the statistical prediction method, whereas a positive d^+ indicates superiority of the clinical method. An effect of zero indicates exactly no difference between the two methods. In addition to d^+, we reported the 95% confidence interval for the effect size. Confidence interval provides the same information as that extracted from significant tests. It permits one to say with 95% confidence (i.e., $\alpha = .05$) that the true effect size falls within its boundaries. If the confidence interval includes zero, the population effect may be zero; one cannot say with confidence that a meaningful difference exists between the two groups. However, if the confidence interval does not include zero, one can conclude that a reliable difference exists between clinical and statistical prediction (e.g., Johnson, 1993).

The data were reduced to one representative effect size per study in most cases. This prevented bias that would result if a single study was overrepresented in the sample (Cooper, 1998; Rosenthal, 1991). For instance, if a study reported more than one statistical or clinical prediction (e.g., brain impairment and lateralization of the impairment; Adams, 1974), an average of the reported judgment accuracy statistic was calculated and transformed into one effect size. Also, if a study reported results from both non–cross-validated and cross-validated statistical prediction schemes, only results from the cross-validated statistical formula were used. This was done to prevent bias in favor of the statistical method, given the possibility of inflated correlations (based on spurious relations) between predictor and criterion variables in non–cross-validated statistical formulas (for more discussion of these issues, see Efron & Gong, 1983). Table 1 notes whether the studies used cross- or non–cross-validated statistical formulas.

Even though one average effect size per study was usually calculated, 18 studies produced more than one effect size (see Table 1). These studies included more than one design characteristic (independent variables) that we hypothesized might influence clinical versus statistical prediction accuracy and reported accuracy statistics for various levels of the independent variable. An example would be a study investigating clinical versus statistical prediction under two conditions. In one condition, the clinicians have access to the statistical prediction scheme, whereas in another condition they do not. In our studies, we extracted two effect sizes. That is, the study's two conditions (with and without access to the statistical formula) were treated as two independent projects. Furthermore, a study was allowed to produce

TABLE 1: Studies Included in Meta-Analysis

Citation	*Prediction*	*Accuracy Statistic Reported*	*Accuracy*		
			Clinical	*Statistical*	d^+
Adams (1974)[1]	Brain impairment	Hit rate	53	52	.02
Adams (1974)[2a]	Brain impairment	Hit rate	53	56	−.06
Alexakos (1966)[1]	Academic performance	Hit rate	39	51	−.24
Alexakos (1966)[2]	Academic performance	Hit rate	39	52	−.27
Astrup (1975)[b]	Psychiatric diagnosis	Hit rate	78	74	.09
Barron (1953)	Psychotherapy outcome	Hit rate	62	73	−.23
Blumetti (1972)	Length of psychotherapy	Hit rate	61	54	.15
Bolton et al. (1968)	Prognosis	Correlation	.35	.48	−.16
Carlin and Hewitt (1990)	Real vs. random MMPI profile	Hit rate	63	95	−.73
Conrad and Satter (1954)	Academic performance	Correlation	.36	.46	−.12
Cooke (1967a)	Psychiatric diagnosis	Hit rate	77	76	.02
Cooke (1967b)[a]	Psychiatric diagnosis	Correlation	.42	.51	−.11
Danet (1965)	Prognosis	Hit rate	64	70	−.13
Devries and Shneidman (1967)[a]	Matching MMPI profiles to persons	Hit rate	75	100	−.81
Dickerson (1958)	Compliance with counseling plan	Hit rate	57	52	.10
Dunham and Meltzer (1946)[b]	Length of hospital stay	Hit rate	38	58	−.40
Evenson, Altman, Sletten, and Cho (1975)	Length of hospital stay	Hit rate	64	71	−.14
Fero (1975)[1]	Prognosis	Correlation	.35	.57	−.29
Fero (1975)[2b]	Prognosis	Correlation	.35	.73	−.57
Gardner et al. (1996)[1b]	Offense / violence	Hit rate	62	74	−.25
Gardner et al. (1996)[2]	Offense or violence	Hit rate	62	71	−.20
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	−.17
Gardner et al. (1996)[3]	Offense or violence	Hit rate	62	70	−.17

Gaudette (1992)[a]	Brain impairment	Correlation	.39	.45	−.07
Goldberg (1965) × 65[d]	Psychiatric diagnosis	Hit rate	62	64	−.04
Goldberg (1970)[1a]	Psychiatric diagnosis	Correlation	.28	.31	−.03
Goldberg (1970)[2]	Psychiatric diagnosis	Correlation	.28	.44	−.18
Goldberg (1970)[3a]	Psychiatric diagnosis	Correlation	.28	.46	−.21
Goldstein, Deysach, and Kleinnecht (1973)	Brain impairment	Hit rate	95	75	.57
Grebstein (1963)[a]	IQ	Correlation	.62	.56	.08
Gustafson, Greist, Stauss, Erdman, and Laughren (1977)[a]	Suicide attempt	Hit rate	65	80	−.33
Halbower (1955)	Prognosis	Correlation	.60	.79	−.21
Hall (1988)[a]	Offense or violence	Hit rate	55	81	−.59
Heaton et al. (1981)[b]	Brain damage	Hit rate	79	72	.14
Holland, Holt, Levi, and Beckett (1983)	Offense or violence	Correlation	.21	.28	−.08
Holt (1958)	Academic or training performance	Correlation	.25	.13	.12
Hovey and Stauffacher (1953)	Personality characteristics	Hit rate	74	63	.23
Johnston and McNeal (1967)	Length of hospital stay	Hit rate	72	72	.00
Kaplan (1962)[1]	Prognosis	Hit rate	58	70	−.25
Kaplan (1962)[2]	Prognosis	Hit rate	66	70	−.09
Kelly and Fiske (1950)	Academic or training performance	Correlation	.28	.29	−.01
Klehr (1949)	Psychiatric diagnosis	Hit rate	57	64	−.14
Kleinmuntz (1967)	Adjustment	Hit rate	66	71	−.10
Klinger and Roth (1965)[b]	Psychiatric diagnosis	Hit rate	94	71	.63
Lefkowitz (1973)	Marital adjustment	Hit rate	54	55	−.03
Leli and Filskov (1981)[1]	Brain impairment	Hit rate	41	62	−.44
Leli and Filskov (1981)[2]	Brain impairment	Hit rate	50	62	−.24
Leli and Filskov (1984)[1a]	Brain impairment	Hit rate	60	83	−.52
Leli and Filskov (1984)[2a]	Brain impairment	Hit rate	71	83	−.30
Lemerond (1977)	Suicide attempt	Hit rate	53	50	.06

(continued)

TABLE 1 (continued)

Citation	*Prediction*	*Accuracy Statistic Reported*	*Accuracy*		
			Clinical	*Statistical*	d[+]
Lewis and MacKinney (1961)	Career satisfaction	Correlation	.12	.21	−.09
Lindsey (1965)	Homosexuality	Hit rate	70	57	.03
Lyle and Quast (1976)	Brain impairment	Hit rate	67	68	−.02
McHugh and Apostolakos (1959)[b]	Academic field	Hit rate	70	46	.50
Meehl (1959)	Psychiatric diagnosis	Hit rate	69	74	−.11
Melton (1952)[1b]	Academic performance	Error rate	.47	.40	−.26
Melton (1952)[2]	Academic performance	Error rate	.45	.35	−.35
Meyer (1973)[1]	Psychiatric diagnosis	Hit rate	65	63	.05
Meyer (1973)[2]	Psychiatric diagnosis	Hit rate	67	63	.08
Miller, Kunce, and Getsinger (1972)[a]	Adjustment / employability	Correlation	.65	.28	.49
Moxley (1970)[1a]	Length of psychotherapy stay	Hit rate	55	67	−.25
Moxley (1970)[2a]	Length of psychotherapy stay	Hit rate	61	67	−.13
Moxley (1970)[3a]	Length of psychotherapy stay	Hit rate	67	67	.00
Moxley (1970)[4a]	Length of psychotherapy stay	Hit rate	69	67	.04
Oskamp (1962) × 16[d]	Prior hospitalization	Hit rate	72	65	.13
Oxman, Rosenberg, Schnurr, and Tucker (1988)[1a]	Psychiatric diagnosis	Hit rate	66	80	−.32
Oxman et al. (1988)[2a]	Psychiatric diagnosis	Hit rate	66	62	.08
Perez (1976)[1a]	Homicidality	Hit rate	50	83	−.77
Perez (1976)[2a]	Homicidality	Hit rate	51	83	−.77
Perez (1976)[3a]	Homicidality	Hit rate	51	83	−.77
Perez (1976)[4a]	Homicidality	Hit rate	58	83	−.61
Popovics (1983)	IQ	Correlation	.20	.19	.01

Sarbin (1942)	Academic performance	Correlation	.56	.60	−.06
Shaffer, Perlin, Schmidt, and Stephens (1974)[1a]	Suicide attempt	Correlation	.40	.40	.00
Shaffer et al. (1974)[2]	Suicide attempt	Correlation	.40	.18	.24
Shaffer et al. (1974)[3b]	Suicide attempt	Correlation	.40	.12	.30
Shagoury and Satz (1969)[a]	Brain impairment	Hit rate	83	87	−.11
Stricker (1967)	Psychiatric diagnosis	Hit rate	70	79	−.21
Szuko and Kleinmuntz (1981)	Lie detection	Correlation	.23	.52	−.32
Taulbee and Sisson (1957)[1]	Psychiatric diagnosis	Hit rate	64	78	−.31
Taulbee and Sisson (1957)[2]	Psychiatric diagnosis	Hit rate	64	63	.02
Thompson (1952)[b]	Juvenile delinquency	Hit rate	65	90	−.60
Walters et al. (1991)[a]	Malingering	Hit rate	59	73	−.30
Watley (1966)	Academic performance	Hit rate	68	75	−.14
Watley and Vance (1964)[a]	Academic performance	Hit rate	70	80	−.25
Webb et al. (1977)[b]	Occupational choice	Hit rate	35	55	−.41
Wedding (1983)[1]	Brain impairment	Hit rate	55	63	−.17
Wedding (1983)[2]	Brain impairment	Hit rate	55	60	−.11
Weinberg (1957)[1a]	Personality characteristics	Correlation	.49	.54	−.15
Weinberg (1957)[2a]	Personality characteristics	Correlation	.40	.54	−.06
Werner, Rose, Yesavage and Seeman (1984)[a]	Offense / violence	Correlation	.14	.40	−.27
Wiggins and Kohen (1971)[1a]	Academic performance	Correlation	.33	.69	−.51
Wiggins and Kohen (1971)[2a]	Academic performance	Correlation	.33	.50	−.20
Wirt (1956)[a]	Prognosis	Hit rate	54	79	−.54
Wittman and Steinberg (1944)[b]	Prognosis	Hit rate	43	78	−.77

NOTE: Hit rate refers to percentage correct; correlation refers to correlation with the criteria; error rate difference was calculated by a t test that was then transformed into d^t. All accuracy statistics are reported raw (untransformed). d^t: Effect size of difference between clinical and statistical prediction accuracy; negative d^t indicates superiority of statistical prediction. Studies producing more than one effect size have each effect identified by superscripts 1 to 4. MMPI = Minnesota Multiphasic Personality Inventory.

a. Studies that did not use cross-validated statistical formulas.

b. Studies that were diagnosed as outliers. These studies were excluded in some analyses.

c. Studies that were only included in the overall effect. The average effect is reported.

d. Average hit rate for Goldberg (1965) and Oskamp (1962) studies.

more than one effect size if it used more than one statistical prediction scheme (e.g., regression formula and test cutoff score).

Despite the potential for bias, the questions addressed by these analyses were essential to understanding what factors influence clinical versus statistical judgment accuracy. Their value thus outweighed the risk of adding bias to the overall results. Both Cooper (1998) and Johnson and Eagly (2000) suggest that this is acceptable under conditions similar to ours.

Nonetheless, two studies (Goldberg, 1965; Oskamp, 1962) produced so many effect sizes that they were treated as separate cases. Oskamp (1962) reported hit rates of 16 different cross-validated statistical formulas compared with the judgment accuracy of the average clinician. Goldberg (1965) reported the prediction accuracy of 65 different statistical methods compared with the average clinician. To prevent an overrepresentation of data from Goldberg and Oskamp, the overall effect size was reported with and without these two studies. They were not included in our analysis of study design characteristics.

Strategy for Data Analyses

The overall effect size was first calculated for 69 studies, including Goldberg (1965) and Oskamp (1962), producing 173 effect sizes. Second, as noted under "Study Selection," the overall effect size was also calculated without the Goldberg and Oskamp studies, producing 92 effect sizes. Third, the overall effect size was calculated using only the 49 studies that compared cross-validated statistical formulas with the clinical method. This last procedure resulted in 60 effect sizes and eliminated artificially inflated accuracy of statistical methods caused by chance associations among variables (e.g., Dawes et al., 1989; Meehl, 1954). Next, the influence of study design characteristics (independent variables) on the overall effect size of these 49 studies was calculated[3] (see Results for description and hypothesized effects). Because heterogeneity was not eliminated when study design characteristics were considered, an outlier analysis was performed to reduce heterogeneity[4] (Johnson & Eagly, 2000). The 41 studies (48 effect sizes) remaining after this analysis represented the most conservative sample of studies and were used to calculate the influence of study design characteristics on the overall effect.

It was necessary to remove 12 outliers (20%) from the 60 effects produced by the use of cross-validated statistical formulas to reach a homogenous set of effects, $Q\ (47) = 65.83$, $p > .05$. The resulting set of 41 studies produced 48 effects and shared a common effect size. They are sufficiently homogenous to be characterized as coming from the same population. Outliers, by contrast, have extreme effects. Their extremity may be because of error or may be because the studies have characteristics that are deviant

relative to most studies in the sample (Hunter & Smith, 1990). According to Hedges (1987) and others, removing up to 20% of studies to reach a homogeneous set of data is not uncommon for meta-analyses of psychological topics. This conservative sample of studies, those that used cross-validated statistical formulas and did not have unrepresentatively extreme effect sizes, was used to interpret differences between clinical and statistical prediction accuracy.

RESULTS

Studies investigated the clinical and statistical judgment accuracy of the following 11 prediction tasks: brain impairment, personality, length of hospital stay or treatment, diagnosis, adjustment or prognosis, violence or offense, IQ, academic performance, if an MMPI profile was real or fictional, suicide attempt, and homosexuality. In all studies, the prediction accuracy of mental health professionals and graduate students in the mental health field were compared with a statistical prediction method. The studies, prediction tasks, and accuracy statistics are listed in Table 1.

Overall Effect Size

Figure 1 presents a distribution of the difference between clinical and statistical prediction accuracy transformed into a weighted effect size (d^+) in stem-and-leaf diagram format for the most conservatively selected set of effect sizes ($n = 48$). As may be seen, effect sizes ranged from .57 in favor of the clinical method to –.73 in favor of the statistical method. Visual inspection reveals a slight skew in favor of the statistical method. Similar results were obtained when outlier effects were included, without Goldberg (1965) and Oskamp (1962). For these 92 effect sizes, effects ranged from .63 to –.81 (see Note 3).

On the basis of suggestions from the literature (Dawes et al., 1989; Garb, 1994; Grove et al., 2000; Grove & Meehl, 1996; Kleinmuntz, 1990; Meehl, 1954; Russell, 1995; Sawyer, 1966; Wiggins, 1981), we expected that statistical prediction would be, in general, more accurate than clinical prediction. Our hypothesis was confirmed. By using the Grove et al. (2000) approach to interpret the relative difference between clinical and statistical prediction accuracy, Figure 1 reveals that of the 48 effect sizes, 25 effects (52%) favored statistical prediction methods, 18 (38%) reported no difference between the two methods, and 5 (10%) favored the clinical method.

Overall effect sizes (both in the form of d^+ and r^+) for the four sets of analyses performed are presented in Table 2. Effect sizes (d^+) ranged from

0	–.8	
0	–.7	
1	–.7	3
0	–.6	
0	–.6	
0	–.5	
0	–.5	
0	–.4	
1	–.4	4
1	–.3	
2	–.3	21
5	–.2	97655
5	–.2	43110
3	–.1	775
8	–.1	44432110
3	–.0	997
4	–.0	3321
6	.0	012223
4	.0	5688
1	.1	2
1	.1	5
2	.2	34
0	.2	
0	.3	
0	.3	
0	.4	
0	.4	
0	.5	
1	.5	7
0	.6	

FIGURE 1. Stem-and-Leaf Diagram for Clinical and Statistical Prediction Differences of Transformed Effect Sizes (d^+) for the Most Conservative Sample of Studies ($N = 48$ Effects)

NOTE: A negative d^+ indicates superiority of statistical method.

–.16, without Goldberg (1965) and Oskamp (1962), to –.12 for studies using cross-validated formulas with outlier studies excluded. As Table 2 also shows, the test for homogeneity of effects, Q, was significant ($p < .001$) in all instances except when outlier effects were removed from the analysis. This indicates that when outlier effects were included, the variability among effect sizes was too great for a reliable interpretation. Therefore, significant heterogeneous effect sizes (reflected by Q values) must be interpreted with caution. By contrast, confidence can be placed in the overall effect size of $d^+ = -.12$ ($r^+ = -.06$) produced by nonoutlying studies with cross-validated formulas. In this case, the effect size is reliable as indicated by a nonsignificant Q. Additionally, the 95% confidence interval did not cross zero, further supporting the reliability of the effect.

TABLE 2: Overall Effect Sizes for Clinical Versus Statistical Prediction

	Mean Weighted Effect Size (d^+)	*95% Confidence Interval*	r^+	*Homogeneity Index* Q
With Goldberg (1965) and Oskamp (1962) ($N = 69$ studies and 173 effects)	−.16	−.18 to −.14	−.08	460.54***
Without Goldberg (1965) and Oskamp (1962) ($N = 67$ studies and 92 effects)	−.15	−.17 to −.13	−.08	470.71***
Cross-validated studies ($N = 49$ studies and 60 effects)	−.14	−.17 to −.12	−.07	371.82***
Cross-validated studies without outliers ($N = 41$ studies and 48 effects)	−.12	−.14 to −.09	−.06	65.84

NOTE: Negative d^+ indicates superiority of statistical method.
***$p < .001$.

Because of these two conditions, we can conclude that statistical prediction methods are, in general, more accurate than clinical prediction methods. This is a small effect by conventional standards (Cohen, 1988), but it is consistent and reliable. When consistency and reliability are established in a meta-analysis, confidence can be placed in the effect as a true effect (i.e., statistical is better than clinical prediction). Another perspective on the meaning of this effect is possible through the use of a binomial effect size display analysis (Rosenthal & Rubin, 1982). This analysis permits the level of improvement achieved with one decision-making method versus another to be determined. In this case, clinical decisions are accurate 47% of the time, $(.50 - r/2) \times 100$, whereas statistical decisions are accurate 53% of the time, $(.50 + r/2) \times 100$. As a result, the likelihood of a successful decision can go up 13% when statistical rather than clinical methods are used.

Study Design Characteristics

The overall effect size of the difference between clinical and statistical methods was analyzed as a function of study design characteristics (i.e., independent variables hypothesized to influence the overall effect size). We used only the most conservative sample of studies. The study design characteristics were fitted to the effect sizes using a procedure analogous to the analysis of variance (Hedges & Olkin, 1985). The between-class effect (Q_B) in this procedure is similar to a main effect in an analysis of variance. The 95% confidence interval and a test of the homogeneity of the effect size within each class (Q_W) were also calculated. The homogeneity of studies within

a class (Q_W) determines the confidence that can be placed in interpreting the effect size (d^+). If studies are heterogeneous within a class (e.g., prediction task), nonrandom variance in d^+ remains in that class (Johnson, 1993).

Results of these analyses are reported in Table 3. Each study design characteristic or independent variable had several levels. For example, studies varied in how much information (predictor cues) they provided to clinicians and the formula on which clinicians rely in making predictions. Some gave the same amount to the clinician and the statistical formula, others gave the clinicians more, others gave the clinician less, and some did not report how much information was provided to either. Before our analysis, we determined that to have sufficient reliability and representativeness, three or more studies would have to examine a design characteristic level to be included in the analysis. Because only two studies gave the clinician less information than the formula, that condition is not included in the overall test of the design characteristic. Within-class effects (Q_W) values are included in Table 3 for information only. Seven studies did not report how much information was given to clinicians and the formula. These seven were also excluded from our analyses.

Type of prediction. Since Meehl's (1954) original publication, the methodology and the nature of the studies comparing clinical and statistical prediction accuracy have been questioned. One particular criticism concerned the unrepresentativeness of the judgment tasks evaluated. It was argued that unusual predictions (e.g., predicting future grade point average; Holt, 1970) did not provide a fair evaluation of the accuracy of clinicians' judgment. In response to this criticism, subsequent studies broadened the scope of the prediction tasks (cf. Rock, Bransford, & Maisto, 1987). We identified 11 prediction tasks and examined six of them (Table 3).

Effect sizes were expected to vary across different prediction tasks. This expectation was confirmed, $Q_B(5) = 12.42$, $p < .05$. Three of the six prediction tasks that met our criteria for analysis had an effect size in which the confidence interval did not include zero. These were predictions of adjustment or prognosis ($d^+ = -.14$), offense or violence ($d^+ = -.17$), and academic performance ($d^+ = -.14$). For the other three prediction tasks, the confidence interval included zero, which indicates that any differences between the clinical and statistical methods were unreliable. Data were insufficient to analyze differences in determining personality type, estimating client IQ, determining whether an MMPI profile was fictitious, predicting suicide attempt, or determining homosexuality. In no instance was the clinical method consistently more accurate than the statistical method.

Data collection setting. Holt (1958, 1970) stated that in many studies, clinicians were asked to make unusual predictions with limited data. These

TABLE 3: Effect Size by Study Design Characteristics, Cross-Validated Studies Without Outliers ($N = 41$ Studies, 48 Effects)

	Between-Class Effect (Q_B)	*n*	*Mean Weighted Effect Size (d^+)*	*95% CI Lower*	*95% CI Higher*	r^+	*Homogeneity Within Class (Q_W)*
Type of prediction	12.42*						
Brain impairment		7	–.05	–.22	.12	–.03	6.33
Hospital or treatment length		3	–.02	–.13	.09	–.01	2.33
Diagnosis		9	–.05	–.12	.01	–.03	7.39
Adjustment or prognosis		11	–.14	–.19	–.10	–.07	6.69
Future offense or violence		4	–.17	–.24	–.11	–.09	1.46
Academic performance		8	–.14	–.18	–.09	–.07	11.30
IQ		(1)	(.01)	(–.36)	(.00)	(.00)	(0.00)
Personality type		(1)	(.23)	(–.18)	(.64)	(.11)	(0.00)
Random MMPI profile		(1)	(–.73)	(–1.24)	(–.22)	(–.34)	(0.00)
Suicide attempt prediction		(2)	(.16)	(–.06)	(.38)	(.07)	(.63)
Homosexuality		(1)	(.27)	(–.24)	(.78)	(.13)	(0.00)
Data collection setting	6.96**						
Clinicians from same setting as data		21	–.14	–.17	–.11	–.07	28.03
Clinicians not from same setting as data		25	–.06	–.11	–.01	–.03	30.37
Not reported		(2)	(–.06)	(–.31)	(.18)	(–.03)	(.29)
Statistical formula type	11.94**						
Linear statistical formula		27	–.15	–.19	–.12	–.08	32.29
Test cutoff score		8	–.10	–.15	–.06	–.05	8.06
Logically constructed rule		11	–.03	–.09	.03	–.02	11.87
Model of clinical judgment		(2)	(–.14)	(–.23)	(–.05)	(–.07)	(2.47)
Amount of information	4.06*						
Same amount for clinicians and formula		24	–.06	–.11	–.01	–.03	28.26
Clinicians have more than formula		15	–.13	–.16	–.09	–.06	25.77*
Clinicians have less than formula		(2)	(–.12)	(–.42)	(.19)	(–.06)	(1.64)
Not reported		(7)	(–.15)	(–.20)	(–.10)	(–.08)	(3.81)

(continued)

TABLE 3 (continued)

	Between-Class Effect (Q_B)	*n*	*Mean Weighted Effect Size* (d^+)	*95% CI Lower*	*95% CI Higher*	r^+	*Homogeneity Within Class* (Q_W)
Information about base rates	4.25						
Base rate provided		8	–.02	–.12	.08	–.01	5.40
Base rate same as natural setting		8	–.11	–.15	–.06	–.05	6.37
Clinicians do not know base rate		27	–.13	–.16	–.09	–.06	45.56*
Not reported		(5)	(–.14)	(–.19)	(–.09)	(–.07)	(3.04)
Availability of statistical formula	2.70						
Available		5	–.14	–.18	–.09	–.07	7.33
Not available		40	–.09	–.12	–.06	–.05	53.34
Not reported		(3)	(–.15)	(–.21)	(–.10)	(–.08)	(.05)
Clinician expertness	2.47						
Experts in the prediction task		7	–.05	–.14	.03	–.03	14.96*
Nonexperts in the prediction task		41	–.12	–.15	–.10	–.06	48.41
Publication source	2.06						
American Psychological Association journal		19	–.11	–.15	–.07	–.05	22.53
Non–American Psychological Association journal		16	–.11	–.15	–.06	–.05	21.85
Dissertation		13	–.15	–.19	–.10	–.07	19.39
Confidence in criterion for accuracy	.12						
Low		12	–.12	–.17	–.08	–.06	12.03
High		36	–.12	–.14	–.08	–.06	53.69

NOTE: A negative d^+ indicates superiority of the statistical method. Values in parentheses were not included in the overall analysis. Either too few studies examined the design characteristic level, or the study did not report how the characteristic was treated. Values are included for information only. CI = confidence interval. MMPI = Minnesota Multiphasic Personality Inventory.

*$p < .05$. **$p < .01$. ***$p < .001$.

conditions, in Holt's opinion, did not adequately represent the clinical approach and invalidated the comparison between clinical and statistical methods. Yet if the data given to clinicians are from the same setting that the clinicians inhabit, then clinicians should be familiar with the predictors. They should know how to use them to form effective judgments. Based on this rationale, we expected that if the clinicians were drawn from the same setting as were the data (e.g., clinicians who worked in the clinic from which the client data were drawn), the difference between clinical and statistical prediction would be less than if the clinicians and the data were not from the same setting.

We found that data collection setting indeed influenced effect size, Q_B (1) = 6.96, $p < .01$, although not as we expected. Contrary to our expectations, when clinicians were not from the same setting as the data on which they based their predictions ($d^+ = -.06$), the effect size was smaller than if the data were from their work setting ($d^+ = -.14$). In neither instance did the effect-size confidence interval include zero. Therefore, even though statistical methods yield more accurate predictions independent of the familiarity of clinicians with the data, the difference between the clinical and the statistical method is smaller when clinicians do not come from the same setting as the data used to make their judgments. Thus—and most unexpected—existing studies indicate that clinicians seem to be more accurate when they are working with less familiar or novel information.

Statistical formula type. Studies have compared several types of statistical formula with the clinical method. We organized the studies on four types of formula: linear statistical models (e.g., regression or discriminant function), logically constructed rules (e.g., Goldberg rule for differential diagnosis of psychosis and neurosis from MMPI), test cutoff scores, and models of clinical judgment (mechanization of clinicians' judgment processes). Even though no specific type of statistical formula was expected to yield the most accurate results, on the basis of Dawes and Corrigan's (1974) hypotheses we assumed that simple linear models would be as accurate as more complex models. We also expected, on the basis of narrative and empirical evidence (e.g., Dawes et al., 1989; Grove et al., 2000), that any of the four types of statistical method would be more accurate than the clinical method.

Our hypotheses were partially supported. The type of statistical formula used affected overall effect size, Q_B (3) = 11.94, $p < .01$. Effect sizes (d^+) ranged from –.15 for linear statistical formulas (i.e., regression and discriminant function analysis) to –.03 for logically constructed rules or signs (e.g., patterned MMPI rules). All categories of statistical formulas, except logically constructed rules, yielded effect sizes in which the confidence interval did not include zero. Logically constructed rules did not differ from

clinical prediction methods. Hence, statistical methods are more accurate than clinical methods but only when purely statistical models are used. Logical rules are unexpectedly no better and no worse than clinical methods. Data were insufficient to compare models mechanizing clinicians' judgment processes to the clinical method.

Amount of information. Yet another controversy surrounding clinical versus statistical prediction concerns the amount of information made available to clinicians and the formula. Holt (1970), for instance, noted that in many studies the only data available to clinicians were quantitative (e.g., an MMPI profile). This, Holt contended, placed the clinical method at a serious disadvantage because "given a chance to show what it can do, it has to have meaningful, qualitative data to work with" (p. 343). In contrast, Dawes et al. (1989) argued that even if clinicians were provided more information than the formulas, regardless of whether the information was qualitative, they will still fail to surpass the statistical method (see also Faust, 1986, 2003). This assumption was supported by Grove et al. (2000). To resolve this controversy, we studied the influence of the amount of information provided to the clinicians and the statistical formula. We hypothesized, in accordance with Dawes et al. (1989), Faust (1986, 2003), and Grove et al. (2000), that the differential accuracy between clinical and statistical prediction would not be affected by the amount of informational cues.

Our findings indicate that the amount of information provided to the clinicians influenced their prediction accuracy, $Q_B(2) = 4.06$, $p < .05$. Increasing the amount of information, however, decreased clinicians' judgment accuracy. More information may thus not be better.

Information about base rates. Humans tend to ignore base rate information when making judgments under ambiguous conditions (Kahneman & Tversky, 1973). Several studies reviewed in this meta-analysis investigated the influence of base rate information on clinicians' predictions. Because base rate information is usually underutilized, we anticipated that effect sizes would be the same when clinicians knew the base rate for the outcome they were predicting (i.e., base rate provided or base rate the same as the natural setting) and when base rate information was not provided.

This was the case. Whether clinicians had base rate information available when making judgments did not matter, $Q_B(2) = 4.25$, $p > .05$. Even so, the difference between the two methods tended to diminish when clinicians had base rate information ($d^+ = -.02$). The range of this small effect includes zero, which suggests no difference between the two methods. The statistical method surpassed the clinical method when the clinicians were

not informed about base rate ($d^+ = -.13$) and when the base rate for the prediction was the same as in the clinician's work setting ($d^+ = -.11$).

Availability of statistical formula. An important question raised in the debate about clinical and statistical prediction is whether clinical prediction can be improved when clinicians are given the statistical formula and its outcome. Because some studies provided clinicians with the statistical formula before they made predictions, we could compare clinical and statistical methods with and without access to the formula. In their narrative review of the literature, both Sines (1970) and Dawes et al. (1989) stated that even though clinicians had access to and were free to use the statistical formulas, their predictions remained less accurate than statistical prediction methods. Therefore, we hypothesized that accessibility to the statistical formula would not influence the differential accuracy of clinical and statistical prediction.

As predicted, clinicians' access to the statistical formulas did not improve their accuracy. With and without access to statistical formulas, clinical prediction remained less accurate than statistical prediction, $Q_B(1) = 2.70$, $p > .05$. In the five studies in which clinicians were provided with the statistical formula, the effect size was –.14, and in the 40 studies in which the clinicians did not have access, the effect size was –.09. In both these instances, the confidence interval did not include zero, providing support for the stability of the effect.

Clinical expertness. Holt (1970) noted that studies often failed to compare comparable clinical and statistical prediction methods. For example, the average judge (not the best judge) was compared with the best statistical formulas (for exceptions, see Goldberg, 1965; Oskamp, 1962). Although few studies compared the best judge with the best formula, clinicians identified as experts in the predictions under study may provide a more competitive contrast with the statistical formula. Spengler et al. (2005) found in a recent meta-analysis a small but reliable increase in clinical judgment accuracy with increased experience. They did not, however, assess the impact of expertness as distinct from experience on judgment accuracy. Drawing on the notion that clinicians vary considerably in their ability to make accurate judgments and that the best clinicians can do well (Holt, 1970), we hypothesized that experts in the prediction task would be more accurate than nonexperts. Likewise, we anticipated that the performance of experts would be similar to or better than that of statistical methods.

Our predictions were not supported, $Q_B(1) = 2.47$, $p > .05$, even though a trend in the hypothesized direction was observed. In seven studies, clinicians

were considered experts in the judgment task. These studies yielded an effect size of –.05, although the confidence interval included zero. In the 41 studies in which judges were not considered experts, d^{+} was larger (–.12) and the confidence interval did not cross zero. Thus, when judgments are made by expert clinicians, the difference between clinical and statistical methods seems to disappear. However, when the clinicians are nonexperts, they are consistently outperformed by statistical formulas.

Publication source. The direction of publication bias is that studies reporting significant results are more often published than studies presenting nonsignificant results (e.g., Greenwald, 1975; Rosnow & Rosenthal, 1989). Because of competition for publication in major journals such as those published by the American Psychological Association, studies in these journals may have larger effects. From this assumption, we anticipated that the differential accuracy of clinical and statistical prediction would be larger in studies retrieved from American Psychological Association journals than from studies retrieved from other sources. This hypothesis was not supported, Q_B (2) = 2.06, $p > .05$.

Confidence in the criterion for accuracy. We assigned a dichotomous rating of high or low to the confidence we placed in the reliability of each study's criterion for accuracy (see Spengler et al., 2005). If we determined that the criteria had suspected unreliability, we rated our confidence in the accuracy criterion as low (e.g., peer or supervisor ratings of successful academic training; Kelly & Fiske, 1950). A high rating was given if use of a relatively reliable and valid criterion for accuracy was reported (e.g., neuroradiological or operative verification of brain damage; Heaton, Grant, Anthony, & Lehman, 1981). We hypothesized larger effect sizes among studies with a highly valid and reliable criterion for accuracy owing to the well-known relation between criterion reliability and a ceiling on predictive validity (for further discussion, see Schmidt & Hunter, 1996). This hypothesis was not confirmed, Q_B (1) = .12, $p > .05$. Furthermore, the effect-size confidence interval did not cross zero. Therefore one may conclude that the greater accuracy of statistical over clinical prediction is not affected by our rating of the quality of the criterion used to determine accuracy.

Other independent variables. Three additional independent variables were identified to clarify the differential accuracy of clinical and statistical prediction. These were the number of prediction tasks performed within a study, the number of clinicians making predictions, and year of publication or completion. While more reliable values were expected for higher numbers of predictions and greater numbers of clinicians, no specific hypotheses

were associated with these variables. Similar to assumptions made by Spengler et al. (2005), we anticipated that as study age becomes more recent, the clinical method would fair better in relation to the statistical. This trend would be because of the increasing attention given through the years to improving the accuracy of clinical judgment (e.g., Arkes, 1981, 1991; Dawes et al., 1989; Faust, 1986; Garb, 1989, 1998; Spengler et al., 1995). Regression analyses (i.e., focused comparison; Rosenthal & Rubin, 1982) were performed to examine the effects of these three independent variables on the overall effect size. None of these continuous variables significantly predicted the effect size ($p > .05$).

DISCUSSION

Our examination of the differential prediction accuracy of clinical and statistical methods from studies completed over a 56-year span shows that in general, statistical prediction methods are somewhat more accurate than the clinical method. This confirms, in most instances, the independent and parallel meta-analytic findings of Grove et al. (2000); it is also in accord with earlier narrative reviews of clinical and statistical prediction (e.g., Dawes et al., 1989; Grove & Meehl, 1996; Meehl, 1954; Sawyer, 1966). Using the most conservative sample of studies, we found an effect size of –.12 favoring the accuracy of statistical over clinical methods. This overall effect is virtually identical to the effect size (.12) of Grove et al. across a wide range of decision contexts. Note that although the signs of effects from the two studies are different, this difference is only because of the opposite coding method used. These effect sizes reflect a 13% increase in accuracy using statistical rather than clinical prediction techniques.

Should a relatively small effect, such as that which we observed, be dismissed as unimportant? We think not. First, partisan arguments have appeared in the literature that strongly favor either clinical or statistical predictions. Our analysis and that of Grove et al. (2000) argue for more temperance on both sides. Although the statistical method is almost always the equal of the clinical method and is often better, the improvement is not overwhelming. Much more research is needed—in particular, programmatic lines of research on statistical prediction—that translates into practical applications for practicing psychologists (e.g., Quinsey et al., 1998). Likewise, supporters of clinical decision making must show how their approach can be improved.

A second reason for not ignoring such a modest effect involves how it is to be used. Consider heart attack prevention. Would not any improvement here be important? The Steering Committee of the Physicians Health Study

Research Group (1988) thought so. They discontinued a randomized double-blind experiment on the use of aspirin to reduce heart attacks after finding a preliminary effect of $r = .034$. This small effect was nonetheless large enough that continuing to give control subjects a placebo would have been unethical. After converting this study's d^+ values (−.12) to r^+ (= −.06), the effect of aspirin in reducing heart attack is half the size of the effect found in the current meta-analysis.

One area in which the statistical method is most clearly superior to the clinical approach is the prediction of violence, $r = -.09$. Out of 1,000 predictions of violence, the statistical method should correctly identify 90 more violent clients than will the clinical method (Rosenthal, 1991). The victims of violence would not consider this effect small. Some predictions are more important; therefore, we recommend that statistical prediction techniques be developed, considered, and used for the most important types of decisions made by counseling psychologists and other mental health professionals (e.g., danger, suicide, and/or parole).

Our study went beyond the work of Grove et al. (2000) by examining how specific aspects of the prediction studies influenced the nature of the difference between clinical and statistical prediction. These study design variables were chosen because they had been hypothesized by other reviewers to explain why clinical and statistical predictions differ in accuracy. We found that in some instances, the difference between clinical and statistical prediction was influenced by study design characteristics. Knowing the type of prediction is important: Predictions of violence or academic performance were much more accurate with statistical techniques, whereas treatment length was predicted equally well by both methods.

We found that being familiar with the prediction setting did not help clinicians do better than statistical methods; in fact, clinicians fared worse. The type of statistical formula was also important. All statistical types, except logically constructed rules, did better than clinicians. Even with the exception, the result was a draw; the clinician did not do better. Some have argued that certain studies have "stacked the deck" against the clinician by providing less information than the formula (e.g., Holt, 1970). We found that when clinicians were given the same or more information than the statistical formula, the formula did better. Information was insufficient to assess those conditions in which clinicians had less information. In contrast, the overall effect size was not influenced by clinicians' access to the statistical formula. Furthermore, the overall effect size was not influenced by the outlet in which the study was published. Finally, trends were observed that future studies should address. Clinicians considered experts in a prediction task did better than nonexperts and did as well as statistical methods.

Also, having base rate information available resulted in clinicians' approaching the prediction accuracy of statistical methods.

Implications for Practice

On the basis of our analyses, we tentatively suggest when and under what conditions counseling psychologists and other mental health practitioners might best use clinical or statistical methods to make accurate predictions about their clients (Westen & Weinberger, 2004). Certain limitations apply, however. Akin to the psychotherapy literature on empirically validated treatments (e.g., Waehler, Kalodner, Wampold, & Lichtenberg, 2000; Wampold, 1997), recommendations about what works for clinical predictions can logically be made based only on those applications that have so far been tested. Likewise, for many hypotheses that we tested, the sample sizes were so small that more research is needed before firm conclusions can be made about those judgments.

Given the convergence between our meta-analysis and the work of Grove et al. (2000), statistical rules ought to be employed when feasible. This is especially true if judgment accuracy is important and errors are costly. This recommendation, however, is made with qualification. First, as was recognized by an American Psychological Association task force on the use of psychological assessment (Meyer et al., 1998), prediction rules for many judgment tasks are scarce. That is, researchers have yet to study a large array of possible applications of statistical prediction techniques.

Second, not all statistical formulas are effective. Examples are Goldberg's rules for predicting neurotic versus psychotic diagnoses from the MMPI. These logically constructed rules were developed from large samples and were extensively cross-validated. A review of 406 samples of patients and nonpatients found that Goldberg's index did not generalize well across samples (Zalewski & Gottesman, 1991). This finding corresponds to our results showing that logically constructed rules, despite being cross-validated, were not more accurate than clinician judgments. Other statistical prediction methods, in contrast, have been found to be successful (e.g., regression formulas for predicting recidivism; Hilton et al., 2004; Quinsey et al., 1998).

Third, counseling psychologists should educate and familiarize themselves with available statistical prediction methods, such as regression formulas, test cutoff scores, and hit rates (for further information, see Anastasi & Urbina, 1996; Crocker & Algina, 1986; Greene, 2000). This is especially true for critical decisions in which false-negative judgments can be costly. Even a small increase in accuracy is important if one is predicting suicide, domestic violence, or postparole adjustment. For example, statistical formulas

exist to predict violent behavior (Hilton et al., 2004; Quinsey et al., 1998), as do cutoff scores and hit rates for improving classification accuracy using several psychological tests (e.g., MMPI/MMPI-2 cutoffs for malingering; Graham, Watts, & Timbrook, 1991; posttraumatic stress disorder; Bury & Bagby, 2002; substance abuse; Rouse et al., 1999; Stein et al., 1999). Ignoring available statistical prediction schemes may do a disservice to clients and could even be unethical when false-negative outcomes carry severe consequences (e.g., Dawes, 2002). Counseling psychologists already rely on nomothetic assessment techniques when making predictions about clients (e.g., Strong Interest Inventory, MMPI-2/MMPI-A, or Millon Multiaxial Clinical Inventory-III). Use of tests and inventories as statistical strategies requires knowledge of models and test and measurement principles to effectively bridge nomothetic findings to the individual situation (for further discussion, see Faust, 1997). Furthermore, counseling psychologists already rely on test cutoff scores, hit rates, and decision trees to aid accurate classification (e.g., the Substance Abuse Subtle Screening Inventory; Miller, 1985). Our findings suggest that counseling psychologists should highly weigh scores from valid and reliable psychological tests in their clinical decision making.

Fourth, when counseling psychologists have familiarized themselves with available statistical formulas and prediction techniques, they should use those formulas and techniques to improve their prediction accuracy. While this is especially true when judgment accuracy is critical, such as predicting future violence and offense, it is also true for predictions of prognosis, psychological adjustment, and academic performance. Several suggestions can be drawn from the literature reviewed here. For instance, in studies examining prognostic predictions (e.g., improvement in psychotherapy; Barron, 1953; Bolton, Butler, & Wright, 1968; Kaplan, 1962; Wirt, 1956; Wittman & Steinberg, 1944), the statistical prediction involved cutoff scores from psychological tests in almost all instances (e.g., MMPI profile within normal limits and high scores on Barron's ego strength scale). Thus on the basis of these findings, counseling psychologists should place substantial weight on valid and reliable instruments when making prognostic predictions about their clients. Furthermore, counseling psychologists should rely on past academic accomplishment and scores from aptitude tests when predicting academic success, especially as it relates to completion of undergraduate training (Alexakos, 1966; Conrad & Satter, 1954; Melton, 1952; Sarbin, 1942; Watley, 1966). We call on research to develop statistical formulas especially relevant to counseling psychology practice to perform these and other tasks. For instance, empirical tests of the relevance of GPA and Graduate Record Examination scores in predicting success in counseling

psychology programs are needed, either alone or in combination with other hypothesized predictors of successful training.

Fifth, counseling psychologists should use available base rate information when making decisions to aid their accuracy. Prevalence data routinely reported in epidemiological studies (e.g., National Comorbidity Survey; Kessler et al., 2003) and reported in *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; *DSM-IV-TR*; American Psychiatric Association, 2000) are useful, as are normative data for specific settings or client type. Prevalence data may aid diagnostic accuracy such that more prevalent diagnoses (higher base rate) are more likely to be accurate than less prevalent diagnoses, given certain symptoms. Spengler et al. (1995) hypothesized that teaching counseling psychologists how to use base rates in their decision making would increase judgment accuracy, and they found support for this in a pre- and posttest control group comparison of a 6-week educational course on clinical decision making (see Spengler & Strohmer, 2001). Further research is needed on the use of base rates for increasing the accuracy of clinical judgments.

Sixth, counseling psychologists should be skeptical about the relative accuracy of their clinical judgments, even when they are working with familiar cases in familiar contexts. Studies show that clinical predictions were worse than statistical predictions in these conditions. These counterintuitive results may be because of clinicians' overconfidence or habituation to the setting (e.g., Arkes, 1981, 1991; Faust, 1986). When working with familiar cases, clinicians may become overly confident in their decision making and use a positive hypothesis-testing strategy. Such strategies serve to confirm hypotheses by recalling supporting information and discounting or forgetting contradictory information. Incorporating statistical methods in decision making involving familiar cases could reduce this bias.

Finally, in several instances, clinical judgment was the equal of statistical methods. Our findings show that clinicians can rely on either their own inferential processes (the clinical method) or on statistical methods when detecting brain impairment from psychological tests, determining personality characteristics, predicting length of hospitalization and treatment, and diagnosing. In these predictions, clinicians were as accurate as the statistical approaches. Two reasons might account for these findings. First, experience with these tasks may have provided clinicians with information regarding validity of relations between client data and outcome (Dawes, 1994; Ericsson & Lehman, 1996; Faust, 1991; Lichtenberg, 1997; Spengler, 1998). In fact, these prediction tasks, in comparison with the tasks in which mechanical methods were more accurate, tend to offer a greater opportunity for immediate feedback to the practicing clinician. Second, these prediction tasks may have more

objective tools available to aid judgment accuracy, such as psychological tests and systematic guidelines (e.g., use of *DSM-IV-TR* and test cutoff score for classification), than other prediction tasks in which statistical methods surpassed the clinical method.

Implications for Training

Counseling psychology has given much attention to the importance of the scientist-practitioner model of training. Epistemological issues in research and practice have been discussed, and the use of empirically validated treatment strategies for specific clients and problems has been encouraged. Our analysis supports the argument that training should also encourage the use of statistical methods to decrease judgment biases and errors. A portion of pre- and postdoctoral training could be used to show trainees how frequently they have the opportunity to use statistical methods of prediction and how these methods could improve their predictions. Test cutoff scores, hit rates, and decision trees are all statistical strategies and can improve clinical predictions. Adequate training in statistics and probability theory has been found to increase judgment accuracy (Nisbett & Ross, 1980), as has education about inferential errors and common heuristics (e.g., Arkes, 1981, 1991; Spengler & Strohmer, 2001). Methods for demystifying statistical prediction should be incorporated into counseling psychology curricula. Furthermore, trainees should be familiarized with the construction of simple regression models to aid judgment accuracy. Some research on the benefits of training counselors in effective decision-making strategies already exists (Berven, 1985; Berven & Scofield, 1980; Falvey & Hebert, 1992; Kurpius, Benhamin, & Morran, 1985; Spengler & Strohmer, 2001). More research is needed on how best to implement statistical decision making and comparing its effectiveness in different contexts before educators and trainers can most effectively introduce statistical models of decision making.

In their proposed scientist-practitioner model for assessment, Spengler et al. (1995) suggested that counseling psychologists collect local clinical data for decisions that might be considered routine; these data would be used to evaluate and improve the effectiveness of their practice. Psychologists could then use these data to mechanize routine decisions. Counseling psychologists could also adjust existing statistical formulas to fit their client type and setting, basing their adjustment on local data. Some of the many decisions that might be assisted by statistical formulas include amount of treatment, school-to-work transition, career choice, and predictions of dangerousness to self or others. Blocher (1987) provided a simple and accessible model for counseling psychologists to learn how to empirically evaluate

their practice. Likewise, Stricker (e.g., Stricker & Trierweiler, 1995) has written extensively about training psychologists as local clinical scientists who collect local practice data to improve their decision making.

Counseling psychologists should be trained in statistical methods for evaluating their effectiveness with individual clients. Examples of promising statistical prediction techniques for clinical practice come from a newer area of psychotherapy research called patient-focused research (Howard, Moras, Bril, Martinovich, & Lutz, 1996). By using statistical techniques, such as probit analysis, survival analysis, and hierarchical linear modeling, an individual client's progress is compared with expected recovery curves to improve clinical decision making. Lambert, Hansen, and Finch (2001) developed a methodology based on a data set of 10,000 cases to identify clients likely to fail in treatment (i.e., signal cases). They used the Outcome Questionnaire-45 (Lambert et al., 1996) to monitor client weekly progress. Providing clinicians with the simple feedback to change the course of their treatment, compared with a no-feedback group, led to statistically and clinically significant improvements in recovery rates for the signal cases. Another area of statistical prediction research relevant to counseling training applications is the use of statistical techniques to measure clinically significant and reliable change (Ogles, Lambert, & Masters, 1996).

Implications for Research

This meta-analysis represents only the second meta-analysis conducted in this area of the literature (cf. Grove et al., 2000). The present findings are not without limitations. The arguments in favor of the small, but reliable, edge of statistical prediction techniques are strong, but we are struck by the limits of these studies. Few programmatic lines of research have accumulated bodies of evidence for specific applications (e.g., clinical versus statistical methods to aid suicide assessment risk). More systematic studies, such as those performed by Goldberg (1965) and Quinsey et al. (1998), are clearly needed. Likewise, more models should be developed for clinical practice. Otherwise, these findings will remain only an academic issue of little practical interest to mental health professionals.

A recent proposal by Katsikopoulos, Machery, Pachur, and Wallin (2004) suggested development of user-friendly models, arguing that statistical models must be context based, simple, and friendly for the clinician. Katsikopolous et al. proposed constructing statistical methods they termed "friendly heuristics." These friendly heuristics are simple guidelines that rely on a limited amount of client data and do not need complex integration. Often, these friendly heuristics mirror the cognitive processes underlying clinical judgment.

Examples include relying on few but important client data to render a prediction (e.g., if X communicates a desire to die and has access to a gun, X is likely to commit suicide). Similarly, test cutoff scores and hit rates reported for many psychological tests are simple and easy to use and therefore friendly. To further aid judgment accuracy and save time, more friendly heuristics for use in clinical settings must be developed and tested.

A common criticism of the body of studies on clinical versus statistical prediction accuracy is lack of ecological validity (e.g., Holt, 1970; Rock et al., 1987; Westen & Weinberger, 2004). This criticism has some merit. In various studies, the clinicians made global dichotomous judgments about events that they seldom encounter and about which they may have limited knowledge or training. More research is clearly needed on predictions common to counseling psychologists that incorporate statistical and clinical prediction methods.

In conclusion, we do not argue against clinical prediction as a decision-making strategy (cf. Hammond, 1996). Counseling psychologists make far too many decisions in which the absolute right answer is not the issue and in which determining statistical decision rules may never be practical (e.g., moment-to-moment decisions; Spengler et al., 1985). For these decisions, the clinical strategy is necessary. However, after reviewing 56 years of research, we conclude that clinical prediction should not be the only method. After being shown by two meta-analyses and several independent analyses to be at least equal and often superior to clinical decision making in counseling psychology and mental health contexts, statistical methods must be one of the strategies of the careful clinician. Quoting Meehl (1954), "We have no right to assume that entering the clinic has resulted in some miraculous mutations and made us singularly free from the ordinary human errors which characterized our psychological ancestors" (p. 28).

NOTES

1. A complete description of the search process can be found in Spengler et al. (2005). To limit the retrieval of studies to a manageable, yet representative, sample, studies that appeared between 1970 and 1996 were included in the search. Electronic databases included PsychInfo, ERIC, Dissertation Abstracts, BRS, MEDLINE, and Social Science Index. Unavailable dissertations and journal articles were purchased, and authors were contacted to obtain material that was difficult to retrieve. After we identified likely clinical judgment studies, forward and backward cross-referencing was conducted until no new studies were obtained. By using this strategy, more than 35,000 articles were identified; 4,617 were coded and 1,135 met our inclusion criteria for the project. We chose this open-ended strategy to maximize the number of studies that would be reviewed. Because our search did not include studies published after 1996, we performed a file drawer analysis. In this analysis, we can project how many studies reporting significantly different effect sizes will be needed to change our overall results. This analysis indicated that to reduce our effect size to zero (within the 95% confidence interval

boundaries), indicating no difference between clinical and statistical prediction, 99 additional studies are needed using cross-validated formulas and producing nonoutlying effect sizes.

2. Hit rates reported as percentage correct were directly transformed into d using the DSTAT program (Johnson, 1993). When hit rates were reported as a correlation of the prediction with the criterion (r) for clinical and statistical prediction, the r values were first converted to Z scores. Differences in Z scores were then transformed into a chi-square value. The chi-square value of Z score differences was transformed into d (Rosenthal, 1991).

3. These results are not reported but are available on request by contacting the first author.

4. Johnson (1993) and Hedges and Olkin (1985) recommend outlier analyses in meta-analyses when moderating variables fail to explain observed heterogeneity of studies. In this procedure, outliers are sequentially removed until the hypothesis of homogeneity cannot be rejected (i.e., the probability of Hedges's Q exceeds .05). In the current meta-analysis, the study design characteristics that we hypothesized would influence the overall effect size (e.g., prediction task) did not reduce the heterogeneity among the studies. This made the interpretation of the effect size difficult because the many studies differed from each other in terms of magnitude and direction of the effects. Therefore to make our findings more interpretable, we performed an outlier analysis.

REFERENCES

*References marked with an asterisk indicate studies included in the meta-analysis.

Adams, K. M. (1974). *Automated clinical interpretation of the neuropsychological battery: An ability-based approach.* Unpublished doctoral dissertation, Wayne State University, Detroit, Michigan.

*Alexakos, C. E. (1966). Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions. *Journal of Educational Psychology, 57,* 297-306.

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

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49,* 323-330.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110,* 486-498.

Anastasi, A., & Urbina, S. (1996). *Psychological testing* (7th ed.). New York: Prentice Hall.

*Astrup, C. A. (1975). Predicted and observed outcome in followed-up functional psychosis. *Biological Psychiatry, 10,* 323-328.

*Barron, F. (1953). Some test correlates of response to psychotherapy. *Journal of Consulting Psychology, 17,* 235-241.

Berven, N. L. (1985). Reliability and validity of standardized case management simulations. *Journal of Counseling Psychology, 32,* 397-409.

Berven, N. L., & Scofield, M. E. (1980). Evaluation of clinical problem-solving skills through standardized case-management simulations. *Journal of Counseling Psychology, 27,* 199-208.

Blocher, D. H. (1987). Process models for professional counseling. In *The professional counselor.* New York: MacMillan.

Blumetti, A. E. (1972). *A test of clinical versus actuarial prediction: A consideration of accuracy and cognitive functioning.* Unpublished doctoral dissertation, University of Florida, Gainesville.

Bolton, B. F., Butler, A. J., & Wright, G. N. (1968). Clinical versus statistical prediction of client feasibility. *Wisconsin Studies in Vocational Rehabilitation* [Monograph VII], University of Wisconsin Regional Rehabilitation Research Institute, Madison, WI.

Bury, A. S., & Bagby, R. M. (2002). The detection of feigned uncoached and coached posttraumatic stress disorder with the MMPI-2 in a sample of workplace accident victims. *Psychological Assessment, 14*, 472-484.

*Carlin, A. S., & Hewitt, P. L. (1990). The discrimination of patient-generated and randomly generated MMPIs. *Journal of Personality Assessment, 54*, 24-29.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conrad, H. S., & Satter, G. A. (1954). *The use of test scores and quality-classification ratings in predicting success in electrician's mates school. Project N-106: Research and development of the Navy's aptitude testing program.* Princeton, NJ: Research and Statistical Laboratory College Entrance Examination Board.

*Cooke, J. K. (1967a). Clinicians' decisions as a basis for deriving actuarial formulae. *Journal of Clinical Psychology, 23*, 232-233.

*Cooke, J. K. (1967b). MMPI in actuarial diagnosis of psychological disturbance among college males. *Journal of Counseling Psychology, 14*, 474-477.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

*Danet, B. N. (1965). Prediction of mental illness in college students on the basis of "nonpsychiatric" MMPI profiles. *Journal of Counseling Psychology, 29*, 577-580.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

Dawes, R. M. (2002). The ethics of using or not using statistical prediction rules in psychological practice and related consulting activities. *Philosophy of Science, 69*, S178-S184.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawes, R. M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

*Devries, A. G., & Shneidman, E. S. (1967). Multiple MMPI profiles of suicidal persons. *Psychological Reports, 21*, 401-405.

*Dickerson, J. H. (1958). *The Biographical Inventory compared with clinical prediction of post counseling behavior of V.A. hospital counselors.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

*Dunham, H. W., & Meltzer, B. N. (1946). Predicting length of hospitalization of mental patients. *American Journal of Sociology, 52*, 123-131.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*, 36-48.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273-305.

*Evenson, R. C., Altman, H., Sletten, I. W., & Cho, D. W. (1975). Accuracy of actuarial and clinical predictions for length of stay and unauthorized absence. *Diseases of the Nervous System, 36*, 250-252.

Falvey, J. E., & Hebert, D. J. (1992). Psychometric study of clinical treatment planning simulation (CTPS) for assessing clinical judgment. *Journal of Mental Health Counseling, 14*, 490-507.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology, Research, and Practice, 17*, 420-430.

Faust, D. (1991). What if we had really listened? Present reflections on altered pasts. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. I. Matters of public interest* (pp. 185-217). Minneapolis: University of Minnesota Press.

Faust, D. (1997). Of science, meta-science, and clinical practice: The generalization of a generalization to a particular. *Journal of Personality Assessment, 68*, 331-354.

Faust, D. (2003). Holistic thinking is not the whole story: Alternative or adjunct approaches for increasing the accuracy of legal evaluations. *Assessment, 10*, 428-411.

Fero, D. D. (1975). *A lens model analysis of the effects of amount of information and mechanical decision making aid on clinical judgment and confidence*. Unpublished doctoral dissertation, Bowling Green State University, Bowling Green, OH.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*, 387-396.

Garb, H. N. (1994). Toward a second generation of statistical prediction rules in psychodiagnosis and personality assessment. *Computers in Human Behavior, 10*, 377-394.

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

*Gardner, W., Lidz, C. W., Mulvay, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Consulting and Clinical Psychology, 64*, 602-609.

*Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs: General and Applied, 79*(9), 1-27.

*Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving clinical inferences. *Psychological Bulletin, 73*, 422-432.

*Goldstein, S. G., Deysach, R. E., & Kleinknecht, R. A. (1973). Effect of experience and amount of information on identification of cerebral impairment. *Journal of Consulting and Clinical Psychology, 41*, 30-34.

Gottfredson, D. M., & Snyder, H. N. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts* (NCJ 209158). Washington, DC: National Center for Juvenile Justice, Office of Juvenile Justice and Delinquency Prevention.

Graham, J. R., Watts, D., & Timbrook, R. E. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment, 57*, 264-277.

*Grebstein, L. C. (1963). Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. *Journal of Consulting Psychology, 27*, 127-132.

Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Boston: Allyn & Bacon.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

*Gustafson, D. H., Greist, J. H., Stauss, F. F., Erdman, H., & Laughren, T. (1977). A probabilistic system for identifying suicide attempters. *Computers and Biomedical Research, 10*, 83-89.

*Halbower, C. C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by the Minnesota Multiphasic Personality Inventory*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

*Hall, G. C. N. (1988). Criminal behavior as a function of clinical and actuarial variables in a sexual offender population. *Journal of Consulting and Clinical Psychology, 56*, 773-775.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable justice*. New York: Oxford University Press.

Hare, R. D. (1991). *The revised psychopathy checklist*. Toronto, Canada: Multi-Health Systems.

Harvey-Cook, J. E., & Taffler, R. J. (2000). Biodata in professional entry-level selection: Statistical scoring of common format applications. *Journal of Occupational and Organizational Psychology, 73*, 103-118.

*Heaton, R. K., Grant, I., Anthony, W. Z., & Lehman, R. A. (1981). A comparison of clinical and automated interpretation of the Halstead-Reitan battery. *Journal of Clinical Neuropsychology, 3*, 121-141.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science: The empirical cumulativeness of research. *American Psychologist, 42*, 443-455.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, C. A., & Lines, K. J. (2004). A brief actuarial assessment for the prediction of wife assault recidivism: The Ontario Domestic Assault Risk Assessment. *Psychological Assessment, 16*, 267-275.

*Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology, 68*, 203-211.

Holt, R. R. (1958). Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology, 56*, 1-12.

Holt, R. R. (1970). Yet another look at clinical and statistical prediction: Or, is clinical psychology worthwhile? *American Psychologist, 25*, 337-349.

*Hovey, H. B., & Stauffacher, J. C. (1953). Intuitive versus objective prediction from a test. *Journal of Clinical Psychology, 9*, 341-351.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z. M., & Lutz, W. (1996). Efficacy, effectiveness, and patient progress. *American Psychologist, 51*, 1059-1064.

Hunter, J. E., & Smith, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literature.* Hillsdale, NJ: Lawrence Erlbaum.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology.* London: Cambridge University Press.

*Johnston, R., & McNeal, B. F. (1967). Statistical versus clinical prediction: Length of neuropsychiatric hospital stay. *Journal of Abnormal Psychology, 72*, 335-340.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 801*, 237-251.

*Kaplan, R. L. (1962). *A comparison of actuarial and clinical predictions of improvement in psychotherapy.* Unpublished doctoral dissertation, University of California, Los Angeles.

Katsikopoulos, K., Machery, E., Pachur, T., & Wallin, A. (2004). *The search for models of clinical judgment: Fast, frugal, and friendly in Paul Meehl's spirit.* Unpublished manuscript. Retrieved October 15, 2004, from http://jeannicod.ccsd.cnrs.fr/documents/disk0/00/00/05/33/ijn_00000533_00/ijn_00000533_00.pdf

Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist, 5*, 395-406.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., et al. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association, 289*, 3095-3106.

*Klehr, R. (1949). Clinical intuition and test scores as a basis for diagnosis. *Journal of Consulting Psychology, 13*, 34-38.

*Kleinmuntz, B. (1967). Sign and seer: Another example. *Journal of Abnormal Psychology, 72*, 163-165.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*, 296-310.

*Klinger, E., & Roth, I. (1965). Diagnosis of schizophrenia by Rorschach patterns. *Journal of Projective Techniques and Personality Assessment, 29*, 323-335.

Kurpius, D. J., Benjamin, D., & Morran, D. K. (1985). Effect of teaching a cognitive strategy on counselor trainee internal dialogue and clinical hypothesis formulation. *Journal of Counseling Psychology, 32*, 262-271.

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*, 159-172.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ 45.2)*. Wilmington, DE: American Professional Credentialing Services.

*Leli, D. A., & Filskov, S. B. (1981). Clinical-actuarial detection and description of brain impairment with the W-B Form 1. *Journal of Clinical Psychology, 37*, 623-629.

*Leli, D. A., & Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology, 40*, 1435-1441.

Lemerond, J. N. (1977). *Suicide prediction for psychiatric patients: A comparison of the MMPI and clinical judgments*. Unpublished doctoral dissertation, Marquette University, Madison, WI.

*Lewis, E. C., & MacKinney, A. C. (1961). Counselor vs. statistical prediction of job satisfaction in engineering. *Journal of Counseling Psychology, 8*, 224-230.

Lefkowitz, M. B. (1973). *Statistical and clinical approaches to the identification of couples at risk in marriage*. Unpublished doctoral dissertation, University of Florida, Gainesville.

Lichtenberg, J. W. (1997). Expertise in counseling psychology: A concept in search of support. *Educational Psychology Review, 9*, 221-238.

*Lindsey, G. R. (1965). Seer versus sign. *Journal of Experimental Research in Personality, 1*, 17-26.

Lyle, O., & Quast, W. (1976). The Bender Gestalt: Use of clinical judgment versus recall scores in prediction of Huntington's disease. *Journal of Consulting and Clinical Psychology, 44*, 229-232.

McHugh, R. B., & Apostolakos, P. C. (1959). Methodology for the comparison of clinical with actuarial predictions. *Psychological Bulletin, 56*, 301-309.

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

Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. *Journal of Counseling Psychology, 6*, 102-109.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370-375.

*Melton, R. S. (1952). *A comparison of clinical and actuarial methods of prediction with an assessment of the relative accuracy of different clinicians*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Meyer, K. (1973). *The effect of training in the accuracy and appropriateness of clinical judgment*. Unpublished doctoral dissertation, Adelphi University, Garden City, NY.

Meyer, J. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., et al. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group* (Part I). Washington, DC: American Psychological Association.

*Miller, D. E., Kunce, J. T., & Getsinger, S. H. (1972). Prediction of job success for clients with hearing loss. *Rehabilitation Counseling Bulletin, 16*, 21-29.

Miller, G. A. (1985). *The Substance Abuse Subtle Screening Inventory (SASSI): Manual*. Bloomington, IN: Spencer Evening World.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Ogles, B., Lambert, M. J., & Masters, K. S. (1996). *Assessing outcome in clinical practice*. Needham Heights, MA: Allyn & Bacon.

*Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs: General and Applied, 76*, 1-27.

Oxman, T. E., Rosenberg, S. D., Schnurr, P. P., & Tucker, G. J. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry, 145*, 464-468.

Pepinsky, H. B., & Pepinsky, N. (1954). *Counseling theory and practice*. New York: Ronald Press.

*Perez, F. I. (1976). Behavioral analysis of clinical judgment. *Perceptual and Motor Skills, 43*, 711-718.

*Popovics, A. J. (1983). Predictive validities of clinical and actuarial scores of the Gesell Incomplete Man Test. *Perceptual and Motor Skills, 56*, 864-866.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

Rock, D. L., Bransford, J. D., & Maisto, S. A. (1987). The study of clinical judgment: An ecological approach. *Clinical Psychology Review, 7*, 645-661.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. (1982). A simple general purpose display of magnitude of experimental effects. *Journal of Educational Psychology, 74*, 166-169.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Rouse, S. V., Butcher, J. N., & Miller, K. B. (1999). Assessment of substance abuse in psychotherapy clients: The effectiveness of the MMPI-2 substance abuse scales. *Psychological Assessment, 11*, 101-107.

*Russell, E. W. (1995). The accuracy of automated and clinical detection of brain damage and lateralization on neuropsychology. *Neuropsychology Review, 5*, 1-68.

Sarbin, T. L. (1942). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology, 48*, 593-602.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178-200.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

Shaffer, J. W., Perlin, S., Schmidt, C. W., & Stephens, J. H. (1974). The prediction of suicide in schizophrenia. *Journal of Nervous and Mental Disease, 150*, 349-355.

Shagoury, P., & Satz, P. (1969). The effect of statistical information on clinical prediction. *Proceedings of the 77th Annual Convention of the American Psychological Association, 4*, 517-518.

Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry, 116*, 129-144.

Spengler, P. M. (1998). Multicultural assessment and a scientist-practitioner model of psychological assessment. *The Counseling Psychologist, 26*, 930-938.

Spengler, P. M., & Strohmer, D. C. (2001, August). *Empirical analyses of a scientist-practitioner model of assessment*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Spengler, P. M., Strohmer, D. M., 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.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2005). *The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy*. Manuscript submitted for publication.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318*, 262-264.

Stein, L. A. R., Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). Using the MMPI-2 to detect substance abuse in an outpatient mental health setting. *Psychological Assessment, 11*, 94-100.

Stricker, G. (1967). Actuarial, naïve clinical, and sophisticated clinical prediction of pathology from figure drawings. *Journal of Consulting Psychology, 31*, 492-494.

Stricker, G., & Trieweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist, 50*, 995-1002.

Sullivan, E., Cirincione, C., Nelson, K., & Wallis, J. (2001). *Classifying inmates for strategic programming*. Washington, DC: National Criminal Justice Service, U.S. Department of Justice.

*Szuko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36*, 488-496.

*Taulbee, E. S., & Sisson, B. D. (1957). Configurational analysis of MMPI profiles of psychiatric groups. *Journal of Consulting Psychology, 21*, 413-417.

*Thompson, R. E. (1952). A validation of the Glueck Social Prediction Scale for proneness to delinquency. *Journal of Criminal Law, Criminology, and Police Science, 43*, 451-470.

Waehler, C. A., Kalodner, C. R., Wampold, B. E., & Lichtenberg, J. W. (2000). Empirically supported treatments (ESTs) in perspective: Implications for counseling psychology training. *The Counseling Psychologist, 28*, 657-671.

*Walters, G. D., White, T. W., & Greene, R. L. (1987). The use of MMPI to identify malingering and exaggeration of psychiatric symptomatology in male prison inmates. *Journal of Consulting and Clinical Psychology, 1*, 111-117.

Wampold, B. E. (1997). Methodological problems in identifying efficacious psychotherapies. *Psychotherapy Research, 7*, 21-43.

Watley, D. J. (1966). Counselor variability in making accurate predictions. *Journal of Counseling Psychology, 13*, 53-62.

*Watley, D. J., & Vance, F. L. (1964). *Clinical versus actuarial prediction of college achievement and leadership activity*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Webb, S. C., Hultgen, D. D., & Craddick, R. A. (1977). Predicting occupational choice by clinical and statistical methods. *Journal of Counseling Psychology, 24*, 98-110.

*Wedding, D. (1983). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, 5*, 49-55.

*Weinberg, G. H. (1957). *Clinical versus statistical prediction with a method of evaluating a clinical tool*. Unpublished doctoral dissertation, Columbia University, New York.

Werner, P. D., Rose, T. L., Yesavage, J. A., & Seeman, K. (1984). Psychiatrists' judgment of dangerousness in patients on an acute care unit. *American Journal of Psychiatry, 141*, 263-266.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*, 595-613.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review, 1*, 3-18.

Wiggins, N., & Kohen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19*, 100-106.

*Wirt, R. D. (1956). Actuarial prediction. *Journal of Consulting Psychology, 20*, 123-124.
*Wittman, M. P., & Steinberg, L. (1944). Follow-up of an objective evaluation of prognosis in dementia praecox and manic-depressive psychosis. *The Elgin Papers, 5*, 216-227.
Zalewski, C. E., & Gottesman, I. I. (1991). (Hu)man versus mean revisited: MMPI group data and psychiatric diagnosis. *Journal of Abnormal Psychology, 100*, 562-568.

Journal of Abnormal Psychology
2006, Vol. 115, No. 2, 192–194

0021-843X/06/$12.00 DOI: 10.1037/0021-843X.115.2.192

Meehl's Contribution to Clinical Versus Statistical Prediction

William M. Grove and Martin Lloyd
University of Minnesota, Twin Cities Campus

Paul E. Meehl's work on the clinical versus statistical prediction controversy is reviewed. His contributions included the following: putting the controversy center stage in applied psychology; clarifying concepts underpinning the debate (especially his crucial distinction between ways of gathering data and ways of combining them) as well as establishing that the controversy was real and not concocted, analyzing clinical inference from both theoretical and probabilistic points of view, and reviewing studies that compared the accuracy of these 2 methods of data combination. Meehl's (1954/1996) conclusion that statistical prediction consistently outperforms clinical judgment has stood up extremely well for half a century. His conceptual analyses have not been significantly improved since he published them in the 1950s and 1960s. His work in this area contains several citation classics, which are part of the working knowledge of all competent applied psychologists today.

Keywords: clinical prediction, actuarial prediction, statistical prediction, clinical judgment

Prediction is a central, indeed nearly ubiquitous, activity of psychologists. Many clinical decisions, such as treatment selection, depend on predictions. Psychologists, or at least applied psychologists, are, therefore, obliged to know as much as possible about how to make good predictions.

Paul E. Meehl saw this clearly. He focused on practical contexts in which predictions must be made immediately, based on then-available information. Meehl argued that in such contexts psychologists must choose between clinical and statistical prediction methods, and that they should use whichever method yields the most accurate predictions in the long run. His argument proceeded as follows:

1. There are various ways of gathering predictive data, such as interviews, direct observation, and psychometric tests. No matter how gathered, such data can be encoded or quantified. Encoded data can then be combined by a professional using trained judgment or mechanically (e.g., by a formula, actuarial table, or computer program). These are mutually exclusive and exhaustive classes of ways to combine data; the relative value of these two classes is a meaningful question, whether the data to be combined come from interviews, Rorschach protocols, Minnesota Multiphasic Personality Inventory-2 (MMPI-2) profiles, or behavior counts.

2. There is no true hybrid of these data combination methods; this point is too often misunderstood or erroneously denied. True, a clinician can be given, as one predictive datum to consider, the output of a statistical formula; a formula can include a variable representing a quantified clinician judgment. However, in the former situation, the final prediction depends on trained judgment, whereas in the latter it does not. Clinician judgments, for which statistical predictions were available as cues, can and should be studied to learn whether they are more accurate than clinical predictions made in the absence of such cues; likewise, mechanical predictions based in part on quantified judgments can be compared with mechanical predictions based solely on nonjudgmental data. The existence of these subtypes of clinical and mechanical prediction, respectively, does not make the clinical–statistical distinction meaningless, arbitrary, or lacking in practical interest.

3. When both clinical and statistical predictions are available for a given individual, they will not always agree, and then one cannot follow both. For a given prediction problem, it will surely be the case that the two data combination methods do not yield precisely the same accuracy of their predictions. Lacking clairvoyant knowledge of how the current case will turn out, the best way to maximize one's predictive accuracy is to use whichever data combination method yields the most accurate predictions in the long run.

4. Which data combination method is most accurate, for a given prediction problem, is a pragmatic question empirically answerable by running appropriately designed studies.

Logical arguments for and against mechanical prediction had been published since the 1920s, but Sarbin's (1944) review, the most comprehensive available when Meehl began work, covered only four studies. Meehl saw Sarbin's theoretical analysis, which argued a priori for the superiority of mechanical prediction, as not doing justice to the potential flexibility of clinical judgments.

William M. Grove and Martin Lloyd, Department of Psychology, University of Minnesota, Twin Cities.

This article is dedicated to the memory of P. E. Meehl—mentor, colleague, and friend. We thank Leslie J. Yonce for her help in completing this article.

Correspondence concerning this article should be addressed to William M. Grove, Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 East River Rd, Minneapolis, MN 55455-0344. E-mail: grove001@umn.edu

(Sarbin postulated that the clinician essentially engages in the same kinds of weighting-and-adding processes used in statistical prediction formulas, but clinicians calculate less reliably and so are less accurate.) Meehl also had serious reservations about antiactuarial arguments published by some psychologists (e.g., Lundberg, 1941). He analyzed the controversy in a very short but powerful 1954 book, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. It became a citation classic, going through seven printings by its original publisher and brought back into print in 1996 (Meehl, 1954/1996).

Key Features of Meehl's Book

Meehl made four major contributions to the clinical–statistical controversy in his book. First, he sharply distinguished data gathering from data combination, focusing on the accuracy of clinical versus mechanical methods for combining data. He identified statistical (actuarial, mechanical, formal, algorithmic) predictions as those requiring no professional judgment; they could be carried out by a clerk or computer. The clinical (informal, impressionistic, intuitive) method comprised all other sorts of predictions.

Meehl's second major advance was a convincing refutation of an often-repeated claim, namely that the clinical–statistical antithesis is artificial, because both methods can be used together. Implicit in this view is the idea that it is not necessary to choose between the two approaches. In 1986, Paul made this argument very pointedly:

> Some critics [of the 1954 book] asked a question that Dr. Holt still asks, and which I confess I am totally unable to understand. Why should Sarbin and Meehl be fomenting this needless controversy? Let me state. . .that Sarbin and I did not artificially concoct a controversy. . .between two methods that complement each other and work together harmoniously. I think this is a ridiculous position when the context is the pragmatic context of decision making. You have two quite different procedures for combining a finite set of information to arrive at a predictive decision. . . . [These two procedures] disagree a sizable fraction of the time. . . .The plain fact is that [a decision maker] cannot act in accordance with both of [two] incompatible predictions. Nobody disputes that it is possible to improve clinicians' practices by informing them of their track records actuarially. Nobody has ever disputed that the actuary would be well advised to listen to clinicians in setting up the set of variables. (Meehl, 1986, p. 372)

A third contribution was Paul's subtle analysis of clinical judgment. He was quite sympathetic to the clinician's potential for creative insight, not surprisingly because he was a practicing psychoanalytic psychotherapist. Six of the book's 10 chapters concerned topics such as "The Special Powers of the Clinician," "The Problem of the Logical Reconstruction of Clinical Activity," and "Remarks on Clinical Intuition." Meehl argued strongly that Sarbin's assertion, in effect that "the clinician is a second-rate substitute for a Hollerith machine" (Meehl, 1954/1996, p. 76), was erroneous.

Paul became identified in the field as a fervent proactuarial psychologist, especially by people who heard or read about his work but did not carefully read it for themselves. In fact, his book thoroughly analyzed inherent (i.e., irremediable) limitations of actuarial prediction accuracy. Paul's used what became known as the "broken-leg case" to explore this issue. It goes like this: We have observed that Professor A quite regularly goes to the movies on Tuesday nights. Our actuarial data support the inference "If it's a Tuesday night, then Pr{Professor A goes to movies} = .9." However, suppose we learn that Professor A broke his leg Tuesday morning; he's in a hip cast that won't fit in a theater seat. Any neurologically intact clinician will *not* say that Pr{goes to movies} = .9; they'll predict that he won't go. This is a "special power of the clinician" that cannot, in principle, be completely duplicated by even the most sophisticated computer program. That's because there are too many distinct, unanticipated factors affecting Professor A's behavior; the researcher cannot gather good actuarial data on all of them so the program can take them into account.

For several reasons, this argument does not prove that clinicians will generally outpredict statistical formulas. First, Meehl noted that broken legs are rare, so the clinicians avoiding error in such cases will not greatly increase their accuracy compared with statistical prediction. Second, the example involves a well-corroborated theory (namely, skeletal mechanics, which tells us how limbs—with or without casts on them—can and cannot be fitted into spaces), which predicts to near certainty that Professor A will not go to the movies. In contrast, psychological theories are very seldom this reliable; the advantage for a psychologist judge will decrease as the theory, which justifies making an exception to the statistical rule, becomes less dependable or exerts less influence on the behavior in question. In sum, broken-leg cases exist. They offer an opportunity for clinicians to be more than "second-rate Hollerith machines." However, the degree to which broken-leg cases actually allow clinicians to outpredict actuarial tables is an empirical matter. Moreover, it is very difficult to discern, in individual cases, whether one is justified in overriding a statistical prediction, when there appears to be an exceptional set of circumstances.

Meehl's chapter 8 reviewed 22 studies comparing clinical and statistical prediction. His box score, strongly favoring statistical prediction, was the fourth major point; many psychologists only know about this part of the book. (McNemar, 1955, pointed out a statistical error in one reviewed study, which Paul failed to catch, correction of which only strengthened the case in favor of statistical prediction.)

Meehl (1986) said he would, in hindsight, change at most 5% of the book. He thought that in 1954 he had overemphasized the advantage that configural (nonlinear, interactive) data-combining judgments might give clinicians over the actuary.

Meehl's Later Work

Meehl got so much right in 1954 that he did not have to publish substantially amended opinions on this subject. He encouraged applying actuarial prediction to clinical assessment in the 1950s. His student Hallbower (1955) developed the first actuarial code book for MMPI interpretation, assigning profiles to classes based on combinations of several high (and, more rarely, low) clinical scale scores. A manual giving probabilistic predictions about individuals having each code type was provided. Meehl discussed this work and its implications in "Wanted—A Good Cookbook." Many subsequent code book studies were published, based on larger samples (see, e.g., Gilberstadt & Duker, 1965; Marks & Seeman, 1963). This has been the most widely accepted application of actuarial prediction in clinical psychology.

Meehl (1965) mentioned 51 studies known to him that compared clinical and statistical prediction, more than twice the size of the 1954 literature. These investigations almost uniformly confirmed Meehl's earlier conclusion: Statistical prediction essentially always worked at least as well, and usually worked better, than

clinical prediction. Meehl's (1986) "Causes and Effects of My Disturbing Little Book," delivered at the 40th anniversary symposium for his book, relates the genesis of Paul's interest in this problem and the trouble he had publishing his 1954 book (it was refused by two publishers because they confidently predicted it would not sell). Dawes, Faust, and Meehl (1989) briefly summarized the literature but gave most attention to reasons why few clinicians changed their practices in the face of ever-increasing, consistent evidence favoring statistical prediction. We surveyed a 10% random sample of American Psychological Association Division 12 (clinical) psychologists to learn how familiar they were with the controversy, their views on the matter, and their clinical practices. Of 183 responders (28% response rate), more than 15% had never heard of the controversy or had merely heard that it existed; only 42% had covered the controversy in detail during their training; 10% had not been taught that there were any available statistical prediction methods, let alone what they were or how to use them, and another 6% had only had the existence of such methods mentioned. These findings, supplemented by other reasons for not using statistical prediction discussed by Dawes et al., emphasize the need to improve training on this issue.

A meta-analytically based box score regarding 136 comparative studies (66 of them accumulated by Meehl) was reported by Grove and Meehl (1996). Of these studies, just eight notably favored clinical prediction; no study favoring the clinician replicated any other study.

Importance of Paul's Work Today

In my opinion, Meehl's dissection of prediction methods into two incompatible kinds (clinical and mechanical–algorithmic), when psychologists operate in specific practical prediction contexts, is a fundamental and invaluable insight. As Paul pointed out, there may well be reasoning processes that clinicians sometimes use that a formula, table, or computer program cannot precisely mimic. However, whether such reasoning actually helps clinicians dependably outperform statistical formulas and computer programs is an empirical question with a clear, convincing answer: No, for prediction domains thus far studied. The burden of proof is now squarely on clinicians' shoulders to show, for new or existing prediction problems, that they can surpass simple statistical methods in accurately predicting human behavior.

It remains important for applied psychologists in general and cognitive psychologists in particular, and not just clinical psychologists, to know Meehl's work on the clinical versus statistical prediction controversy. Industrial–organizational psychologists (including nonclinical applied psychologists in the military) are frequently faced with personnel selection problems that involve predicting future behaviors (e.g., work output, time lost from work, dishonest acts, advancement in the organization, being a good team player) from interviews, questionnaires, biographical data, and other data. Formally, this is exactly the same kind of data-combining problem faced by clinicians. Moreover, quite a few of the relevant studies in the literature (e.g., the meta-analysis) actually involve work-related behaviors. Similarly, educational and school psychologists have often been asked to predict who will succeed in schooling or training based on previous performances (e.g., grade point), test scores, and other information (e.g., recommendation letters). They are also frequently used to help select students deemed in need of individualized tutelage or nonmainstream education because of predicted failure in regular schooling; this typically involves integrating past performances, cognitive measures, and observed classroom behavior. Cognitive psychologists interested in human judgment, and especially those studying heuristics and biases, will find Meehl's work seminal. He thought long and hard not just about whether statistical formulas do or do not outpredict human judges but also about cognitive processes tending to help clinicians to perform well (e.g., recognizing novel patterns) or poorly (e.g., ignoring base rates, nonoptimal data combining procedures, inconsistent applications of a combining procedure).

Although Meehl's writings are filled with allusions to philosophy and logic, they are actually quite accessible to those who will read them with care. Very few other 50-year-old books in psychology are still being as frequently cited as Meehl's *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (1954/1996). One likely reason for this is that Paul was perhaps the most entertaining writer in clinical psychology, with a uniquely informal style (especially noticeable in articles written during the latter part of his career). His intellectual power bubbles up in nearly every paragraph, creating an exciting learning experience for the reader. His colloquial style and interesting anecdotes increase the attractiveness and accessibility of his work. The pleasure, as well as enlightenment, one can obtain from reading Meehl's work on clinical versus statistical prediction commend his work to a broad audience of behavioral scientists.

References

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1774.

Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation.* Philadelphia: Saunders.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Hallbower, C. C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by MMPI.* Unpublished doctoral dissertation, University of Minnesota.

Lundberg, G. A. (1941). Case studies versus statistical methods: An issue based on misunderstanding. *Sociometry, 4,* 379–383.

Marks, P. A., & Seeman, W. (1963). *Actuarial description of abnormal personality.* Baltimore, MD: Williams & Wilkins.

McNemar, Q. (1955). Review of clinical versus statistical prediction. *American Journal of Psychology, 68,* 510.

Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality, 1,* 27–32.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375.

Meehl, P. E. (1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Northvale, NJ: Jason Aronson. (Original work published 1954)

Sarbin, T. R. (1944). The logic of prediction in psychology. *Psychological Review, 51,* 210–228.

Received February 18, 2004
Revision received March 8, 2004
Accepted March 16, 2004 ■

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

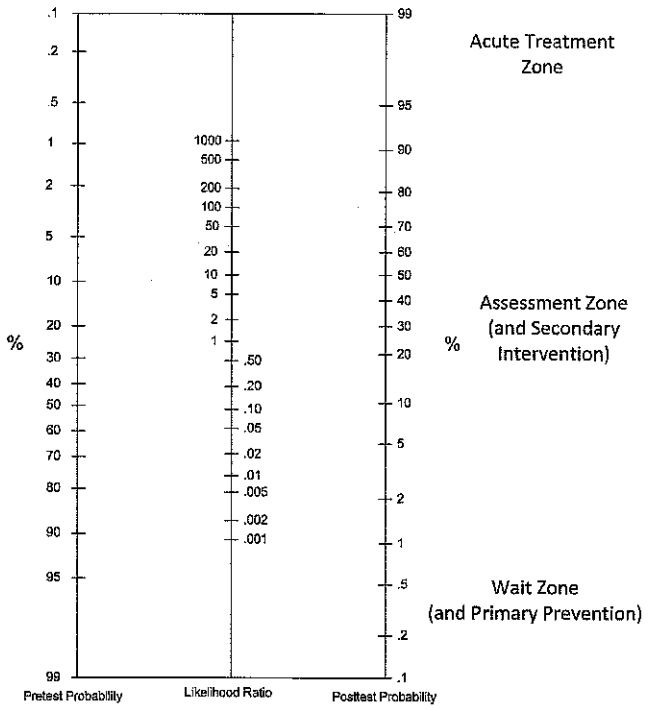


FIGURE 1 Probability nomogram for combining probability with likelihood ratios. *Note:* Straus et al. (2011) provided the rationale and examples of using the nomogram. Jenkins et al. (2011) illustrated using it with a case of possible pediatric bipolar disorder, and Frazier and Youngstrom (2006) with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

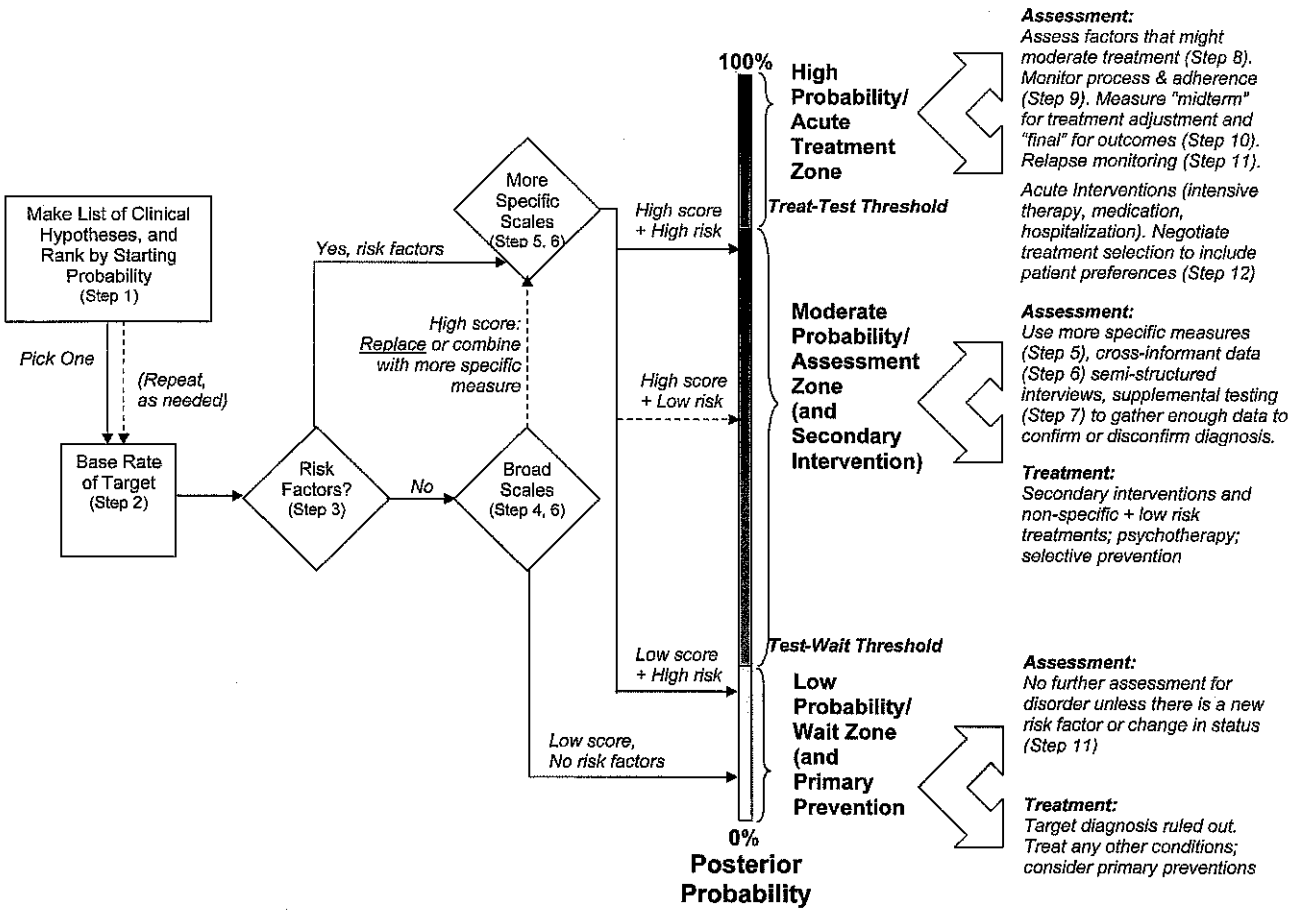


FIGURE 2 Mapping assessment results onto clinical decision making.

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.

Reference to Psychological Testing: History, Principles, and Applications (7th ed.). Boston, MA: Pearson.

CHAPTER 1
Applications and Consequences of Psychological Testing

TOPIC 1A The Nature and Uses of Psychological Testing

1.1 The Consequences of Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec1#ch01lev1sec1)

Case Exhibit 1.1 True-Life Vignettes of Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec1#ch01exh1)

1.2 Definition of a Test (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec2#ch01lev1sec2)

1.3 Further Distinctions in Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec3#ch01lev1sec3)

1.4 Types of Tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec4#ch01lev1sec4)

1.5 Uses of Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec5#ch01lev1sec5)

1.6 Factors Influencing the Soundness of Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec6#ch01lev1sec6)

1.7 Standardized Procedures in Test Administration (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec7#ch01lev1sec7)

1.8 Desirable Procedures of Test Administration (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec8#ch01lev1sec8)

1.9 Influence of the Examiner (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec9#ch01lev1sec9)

1.10 Background and Motivation of the Examinee (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec10#ch01lev1sec10)

If you ask average citizens "What do you know about psychological tests?" they might mention something about intelligence tests, inkblots, and true-false inventories such as the widely familiar MMPI. Most likely, their understanding of tests will focus on quantifying intelligence and detecting personality problems, as this is the common view of how tests are used in our society. Certainly, there is more than a grain of truth to this common view: Measures of personality and intelligence are still the essential mainstays of psychological testing. However, modern test developers have produced many other kinds of tests for diverse and imaginative purposes that even the early pioneers of testing could not have anticipated. The purpose of this chapter is to discuss the varied applications of psychological testing and also to review the ethical and social consequences of this enterprise.

The chapter begins with a panoramic survey of psychological tests and their often surprising applications. In **Topic 1A** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01#ch01box1), The Nature and Uses of Psychological Testing, we summarize the different types and varied applications of modern tests. We also introduce the reader to a host of factors that can influence the soundness of testing such as adherence to standardized procedures, establishment of rapport, and the motivation of the examinee to deceive. In **Topic 1B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec10#ch01box2), Ethical and Social Implications of Testing, we further develop the theme that testing is a consequential endeavor. In this topic, we survey professional guidelines that impact testing and review the influence of cultural background on test results.

1.1 THE CONSEQUENCES OF TESTING

From birth to old age, we encounter tests at almost every turning point in life. The baby's first test conducted immediately after birth is the Apgar test, a quick, multivariate assessment of heart rate, respiration, muscle tone, reflex irritability, and color. The total Apgar score (0 to 10) helps determine the need for any immediate medical attention. Later, a toddler who previously received a low Apgar score might be a candidate for developmental disability assessment. The preschool child may take school-readiness tests. Once a school career begins, each student endures hundreds, perhaps thousands, of academic tests before graduation—not to mention possible tests for learning disability, giftedness, vocational interest, and college admission. After graduation, adults may face tests for job entry, driver's license, security clearance, personality function, marital compatibility, developmental disability, brain dysfunction—the list is nearly endless. Some persons even encounter one final indignity in the frailness of their later years: a test to determine their competency to manage financial affairs.

Tests are used in almost every nation on earth for counseling, selection, and placement. Testing occurs in settings as diverse as schools, civil service, industry, medical clinics, and counseling centers. Most persons have taken dozens of tests and thought nothing of it. Yet, by the time the typical individual reaches retirement age, it is likely that psychological test results will have helped to shape his or her destiny. The deflection of the life course by psychological test results might be subtle, such as when a prospective mathematician qualifies for an accelerated calculus course based on tenth-grade achievement scores. More commonly, psychological test results alter individual destiny in profound ways. Whether a person is admitted to one college and not another, offered one job but refused a second, diagnosed as depressed or not—all such determinations rest, at least in part, on the meaning of test results as interpreted by persons in authority. Put simply, psychological test results change lives. For this reason it is prudent—indeed, almost mandatory—that students of psychology learn about the contemporary uses and occasional abuses of testing. In **Case Exhibit 1.1** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec1#ch01exh1)** , the life-altering aftermath of psychological testing is illustrated by means of several true case history examples.

CASE EXHIBIT 1.1

True-Life Vignettes of Testing

The influence of psychological testing is best illustrated by example. Consider these brief vignettes:

- A shy, withdrawn 7-year-old girl is administered an IQ test by a school psychologist. Her score is phenomenally higher than the teacher expected. The student is admitted to a gifted and talented program where she blossoms into a self-confident and gregarious scholar.
- Three children in a family living near a lead smelter are exposed to the toxic effects of lead dust and suffer neurological damage. Based in part on psychological test results that demonstrate impaired intelligence and shortened attention span in the children, the family receives an $8 million settlement from the company that owns the smelter.
- A candidate for a position as police officer is administered a personality inventory as part of the selection process. The test indicates that the candidate tends to act before thinking and resists supervision from authority figures. Even though he has excellent training and impresses the interviewers, the candidate does not receive a job offer.
- A student, unsure of what career to pursue, takes a vocational interest inventory. The test indicates that she would like the work of a pharmacist. She signs up for a prepharmacy curriculum but finds the classes to be both difficult and boring. After three years, she abandons pharmacy for a major in dance, frustrated that she still faces three more years of college to earn a degree.

These cases demonstrate that test results impact individual lives and the collective social fabric in powerful and far-reaching ways. In the first story about the hidden talent of a 7-year-old girl, cognitive test results changed her life trajectory for the better. In the second case involving the tragic saga of children exposed to lead poisoning, the test data helped redress a social injustice. In the third situation—the impulsive candidate for police officer—personality test results likely served the public interest by tipping the balance against a questionable applicant. But test results do not always provide a positive conclusion. In the last case mentioned above, a young student wasted time and money following the seemingly flawed guidance of a well-known vocational inventory.

The idea of a test is thus a pervasive element of our culture, a feature we take for granted. However, the layperson's notion of a test does not necessarily coincide with the more restrictive view held by psychometricians. A **psychometrician** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss256)** is a specialist in psychology or education who develops and evaluates psychological tests. Because of widespread misunderstandings about the nature of tests, it is fitting that we begin this topic with a fundamental question, one that defines the scope of the entire book: What is a test?

1.2 DEFINITION OF A TEST

A **test** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss326) is a standardized procedure for sampling behavior and describing it with categories or scores. In addition, most tests have norms or standards by which the results can be used to predict other, more important behaviors. We elaborate these characteristics in the sections that follow, but first it is instructive to portray the scope of the definition. Included in this view are traditional tests such as personality questionnaires and intelligence tests, but the definition also subsumes diverse procedures that the reader might not recognize as tests. For example, all of the following could be tests according to the definition used in this book: a checklist for rating the social skills of a youth with mental retardation; a nontimed measure of mastery in adding pairs of three-digit numbers; microcomputer appraisals of reaction time; and even situational tests such as observing an individual working on a group task with two "helpers" who are obstructive and uncooperative.

In sum, tests are enormously varied in their formats and applications. Nonetheless, most tests possess these defining features:

- Standardized procedure
- Behavior sample
- Scores or categories
- Norms or standards
- Prediction of nontest behavior

In the sections that follow, we examine each of these characteristics in more detail. The portrait that we draw pertains especially to norm-referenced tests—tests that use a well-defined population of persons for their interpretive framework. However, the defining characteristics of a test differ slightly for the special case of criterion-referenced tests—tests that measure what a person can do rather than comparing results to the performance levels of others. For this reason, we provide a separate discussion of criterion-referenced tests.

Standardized procedure (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss311) is an essential feature of any psychological test. A test is considered to be *standardized* if the procedures for administering it are uniform from one examiner and setting to another. Of course, standardization depends to some extent on the competence of the examiner. Even the best test can be rendered useless by a careless, poorly trained, or ill-informed tester, as the reader will discover later in this topic. However, most examiners are competent. Standardization, therefore, rests largely on the directions for administration found in the instructional manual that typically accompanies a test.

The formulation of directions is an essential step in the standardization of a test. In order to guarantee uniform administration procedures, the test developer must provide comparable stimulus materials to all testers, specify with considerable precision the oral instructions for each item or subtest, and advise the examiner how to handle a wide range of queries from the examinee.

To illustrate these points, consider the number of different ways a test developer might approach the assessment of *digit span*—the maximum number of orally presented digits a subject can recall from memory. An unstandardized test of digit span might merely suggest that the examiner orally present increasingly long series of numbers until the subject fails. The number of digits in the longest series recalled would then be the subject's digit span. Most readers can discern that such a loosely defined test will lack uniformity from one examiner to another. If the tester is free to improvise any series of digits, what is to prevent him or her from presenting, with the familiar inflection of a television announcer, "1-800-325-3535"? Such a series would be far easier to recall than a more random set, such as, "7-2-8-1-9-4-6-3-7-4-2." The speed of presentation would also crucially affect the uniformity of a digit span test. For purposes of standardization, it is essential that every examiner present each series at a constant rate, for example, one digit per second. Finally, the examiner needs to know how to react to unexpected responses such as a subject asking, "Could you repeat that again?" For obvious reasons, the usual advice is "No."

A psychological test is also a limited sample of behavior. Neither the subject nor the examiner has sufficient time for truly comprehensive testing, even when the test is targeted to a well-defined and finite behavior domain. Thus, practical constraints dictate that a test is only a sample of behavior. Yet, the sample of behavior is of interest only insofar as it permits the examiner to make inferences about the total domain of relevant behaviors. For example, the purpose of a vocabulary test is to determine the examinee's entire word stock by requesting definitions of a very small but carefully selected sample of words. Whether the subject can define the particular 35 words from a vocabulary subtest (e.g., on the Wechsler Adult Intelligence Scale-IV, or the WAIS-IV) is of little direct consequence. But the indirect meaning of such results is of great import because it signals the examinee's general knowledge of vocabulary.

An interesting point—and one little understood by the lay public—is that the test items need not resemble the behaviors that the test is attempting to predict. The essential characteristic of a good test is that it permits the examiner to predict other behaviors—not that it mirrors the to-be-predicted behaviors. If answering "true" to the question "I drink a lot of water" happens to help predict depression, then this seemingly unrelated question is a useful index of depression. Thus, the reader will note that successful prediction is an empirical question answered by appropriate research. While most tests do sample directly from the domain of behaviors they hope to predict, this is not a psychometric requirement.

A psychological test must also permit the derivation of scores or categories. Thorndike (**1918** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1636)) expressed the essential axiom of testing in his famous assertion, "Whatever exists at all exists in some amount." McCall (**1939** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1081)) went a step further, declaring, "Anything that exists in amount can be measured." Testing strives to be a form of measurement akin to procedures in the physical sciences whereby numbers represent abstract dimensions such as weight or temperature. Every test furnishes one or more scores or provides evidence that a person belongs to one category and not another. In short, psychological testing sums up performance in numbers or classifications.

The implicit assumption of the psychometric viewpoint is that tests measure individual differences in traits or characteristics that exist in some vague sense of the word. In most cases, all people are assumed to possess the trait or characteristic being measured, albeit in different amounts. The purpose of the testing is to estimate the amount of the trait or quality possessed by an individual.

In this context, two cautions are worth mentioning. First, every test score will always reflect some degree of measurement error. The imprecision of testing is simply unavoidable: Tests must rely on an external sample of behavior to estimate an unobservable and, therefore, inferred characteristic. Psychometricians often express this fundamental point with an equation:

$$X = T + e$$

where X is the observed score, T is the true score, and e is a positive or negative error component. The best that a test developer can do is make e very small. It can never be completely eliminated, nor can its exact impact be known in the individual case. We discuss the concept of measurement error in **Topic 3B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec5#ch03box2), Concepts of Reliability.

The second caution is that test consumers must be wary of reifying the characteristic being measured. Test results do not represent a *thing* with physical reality. Typically, they portray an abstraction that has been shown to be useful in predicting nontest behaviors. For example, in discussing a person's IQ, psychologists are referring to an abstraction that has no direct, material existence but that is, nonetheless, useful in predicting school achievement and other outcomes.

A psychological test must also possess norms or standards. An examinee's test score is usually interpreted by comparing it with the scores obtained by others on the same test. For this purpose, test developers typically provide **norms** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss223) —a summary of test results for a large and representative group of subjects (**Petersen, Kolen, & Hoover, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1286)). The norm group is referred to as the standardization sample.

The selection and testing of the **standardization sample** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss310) is crucial to the usefulness of a test. This group must be representative of the population for whom the test is intended or else it is not possible to determine an examinee's relative standing. In the extreme case when norms are not provided, the examiner can make no use of the test results at all. An exception to this point occurs in the case of criterion-referenced tests, discussed later.

Norms not only establish an average performance but also serve to indicate the frequency with which different high and low scores are obtained. Thus, norms allow the tester to determine the degree to which a score deviates from expectations. Such information can be very important in predicting the nontest behavior of the examinee. Norms are of such overriding importance in test interpretation that we consider them at length in a separate section later in this text.

Finally, tests are not ends in themselves. In general, the ultimate purpose of a test is to predict additional behaviors, other than those directly sampled by the test. Thus, the tester may have more interest in the nontest behaviors predicted by the test than in the test responses per se. Perhaps a concrete example will clarify this point. Suppose an examiner administers an inkblot test to a patient in a psychiatric hospital. Assume that the patient responds to one inkblot by describing it as "eyes peering out." Based on established norms, the examiner might then predict that the subject will be highly suspicious and a poor risk for individual psychotherapy. The purpose of the testing is to arrive at this and similar predictions—not to determine whether the subject perceives eyes staring out from the blots.

The ability of a test to predict nontest behavior is determined by an extensive body of validational research, most of which is conducted after the test is released. But there are no guarantees in the world of psychometric research. It is not unusual for a test developer to publish a promising test, only to read years later that other researchers find it deficient. There is a lesson here for test consumers: The fact that a test exists and purports to measure a certain characteristic is no guarantee of truth in advertising. A test may have a fancy title, precise instructions, elaborate norms, attractive packaging, and preliminary findings—but if in the dispassionate study of independent researchers the test fails to predict appropriate nontest behaviors, then it is useless.

1.3 FURTHER DISTINCTIONS IN TESTING

The chief features of a test previously outlined apply especially to norm-referenced tests, which constitute the vast majority of tests in use. In a **norm-referenced test** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss219), the performance of each examinee is interpreted in reference to a relevant standardization sample (**Petersen, Kolen, & Hoover, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1286)). However, these features are less relevant in the special case of criterion-referenced tests, since these instruments suspend the need for comparing the individual examinee with a reference group. In a **criterion-referenced test** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss80), the objective is to determine where the examinee stands with respect to very tightly defined educational objectives (**Berk, 1984** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib148)). For example, one part of an arithmetic test for 10-year-olds might measure the accuracy level in adding pairs of two-digit numbers. In an untimed test of 20 such problems, accuracy should be nearly perfect. For this kind of test, it really does not matter how the individual examinee compares to others of the same age. What matters is whether the examinee meets an appropriate, specified criterion—for example, 95 percent accuracy. Because there is no comparison to the normative performance of others, this kind of measurement tool is aptly designated a criterion-referenced test. The important distinction here is that, unlike norm-referenced tests, criterion-referenced tests can be meaningfully interpreted without reference to norms. We discuss criterion-referenced tests in more detail in **Topic 3A** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03#ch03box1), Norms and Test Standardization.

Another important distinction is between testing and assessment, which are often considered equivalent. However, they do not mean exactly the same thing. *Assessment* is a more comprehensive term, referring to the entire process of compiling information about a person and using it to make inferences about characteristics and to predict behavior. **Assessment** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss17) can be defined as appraising or estimating the magnitude of one or more attributes in a person. The assessment of human characteristics involves observations, interviews, checklists, inventories, projectives, and other psychological tests. In sum, tests represent only one source of information used in the assessment process. In assessment, the examiner must compare and combine data from different sources. This is an inherently subjective process that requires the examiner to sort out conflicting information and make predictions based on a complex gestalt of data.

The term *assessment* was invented during World War II (WWII) to describe a program to select men for secret service assignment in the Office of Strategic Services (**OSS Assessment Staff, 1948** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1259)). The OSS staff of psychologists and psychiatrists amassed a colossal amount of information on candidates during four grueling days of written tests, interviews, and personality tests. In addition, the assessment process included a variety of real-life situational tests based on the realization that there was a difference between know-how and can-do:

> We made the candidates actually attempt the tasks with their muscles or spoken words, rather than merely indicate on paper how the tasks could be done. We were prompted to introduce realistic tests of ability by such findings as this: that men who earn a high score in Mechanical Comprehension, a paper-and-pencil test, may be below average when it comes to solving mechanical problems with their hands. (**OSS Assessment Staff, 1948** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1259), pp. 41–42)

The situational tests included group tasks of transporting equipment across a raging brook and scaling a 10-foot-high wall, as well as individual scrutiny of the ability to survive a realistic interrogation and to command two uncooperative subordinates in a construction task.

On the basis of the behavioral observations and test results, the OSS staff rated the candidates on dozens of specific traits in such broad categories as leadership, social relations, emotional stability, effective intelligence, and physical ability. These ratings served as the basis for selecting OSS personnel.

1.4 TYPES OF TESTS

Tests can be broadly grouped into two camps: group tests versus individual tests. **Group tests** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss140) are largely pencil-and-paper measures suitable to the testing of large groups of persons at the same time. **Individual tests** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss155) are instruments that by their design and purpose must be administered one on one. An important advantage of individual tests is that the examiner can gauge the level of motivation of the subject and assess the relevance of other factors (e.g., impulsiveness or anxiety) on the test results.

For convenience, we will sort tests into the eight categories depicted in **Table 1.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec4#ch01tab1) . Each of the categories contains norm-referenced, criterion-referenced, individual, and group tests. The reader will note that any typology of tests is a purely arbitrary determination. For example, we could argue for yet another dichotomy: tests that seek to measure maximum performance (e.g., an intelligence test) versus tests that seek to gauge a typical response (e.g., a personality inventory).

In a narrow sense, there are hundreds—perhaps thousands—of different kinds of tests, each measuring a slightly different aspect of the individual. For example, even two tests of intelligence might be arguably different types of measures. One test might reveal the assumption that intelligence is a biological construct best measured through brain waves, whereas another might be rooted in the traditional view that intelligence is exhibited in the capacity to learn acculturated skills such as vocabulary. Lumping both measures under the category of *intelligence tests* is certainly an oversimplification, but nonetheless a useful starting point.

TABLE 1.1 The Main Types of Psychological Tests

Intelligence Tests: Measure an individual's ability in relatively global areas such as verbal comprehension, perceptual organization, or reasoning and thereby help determine potential for scholastic work or certain occupations.
Aptitude Tests: Measure the capability for a relatively specific task or type of skill; aptitude tests are, in effect, a narrow form of ability testing.
Achievement Tests: Measure a person's degree of learning, success, or accomplishment in a subject or task.
Creativity Tests: Assess novel, original thinking and the capacity to find unusual or unexpected solutions, especially for vaguely defined problems.
Personality Tests: Measure the traits, qualities, or behaviors that determine a person's individuality; such tests include checklists, inventories, and projective techniques.
Interest Inventories: Measure an individual's preference for certain activities or topics and thereby help determine occupational choice.
Behavioral Procedures: Objectively describe and count the frequency of a behavior, identifying the antecedents and consequences of the behavior.
Neuropsychological Tests: Measure cognitive, sensory, perceptual, and motor performance to determine the extent, locus, and behavioral consequences of brain damage.

Intelligence tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss162) were originally designed to sample a broad assortment of skills in order to estimate the individual's general intellectual level. The Binet-Simon scales were successful, in part, because they incorporated heterogeneous tasks, including word definitions, memory for designs, comprehension questions, and spatial visualization tasks. The group intelligence tests that blossomed with such profusion during and after WWII also tested diverse abilities—witness the Army Alpha with its eight different sections measuring practical judgment, information, arithmetic, and reasoning, among other skills.

Modern intelligence tests also emulate this historically established pattern by sampling a wide variety of proficiencies deemed important in our culture. In general, the term *intelligence test* refers to a test that yields an overall summary score based on results from a heterogeneous sample of items. Of course, such a test might also provide a profile of subtest scores as well, but it is the overall score that generally attracts the most attention.

Aptitude tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss15) measure one or more clearly defined and relatively homogeneous segments of ability. Such tests come in two varieties: single aptitude tests and multiple aptitude test batteries. A single aptitude test appraises, obviously, only one ability, whereas a multiple aptitude test battery provides a profile of scores for a number of aptitudes.

Aptitude tests are often used to predict success in an occupation, training course, or educational endeavor. For example, the Seashore Measures of Musical Talents (**Seashore, 1938** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1478)), a series of tests covering pitch, loudness, rhythm, time, timbre, and tonal memory, can be used to identify children with potential talent in music. Specialized aptitude tests also exist for the assessment of clerical skills, mechanical abilities, manual dexterity, and artistic ability.

The most common use of aptitude tests is to determine college admissions. Most every college student is familiar with the SAT (Scholastic Assessment Test, previously called the Scholastic Aptitude Test) of the College Entrance Examination Board. This test contains a Verbal section stressing word knowledge and reading comprehension; a Mathematics section stressing algebra, geometry, and insightful reasoning; and a Writing section. In effect, colleges that require certain minimum scores on the SAT for admission are using the test to predict academic success.

Achievement tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss02) measure a person's degree of learning, success, or accomplishment in a subject matter. The implicit assumption of most achievement tests is that the schools have taught the subject matter directly. The purpose of the test is then to determine how much of the material the subject has absorbed or mastered. Achievement tests commonly have several subtests, such as reading, mathematics, language, science, and social studies.

The distinction between aptitude and achievement tests is more a matter of use than content (**Gregory, 1994a** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib646)). In fact, any test can be an aptitude test to the extent that it helps predict future performance. Likewise, any test can be an achievement test insofar as it reflects how much the subject has learned. In practice, then, the distinction between these two kinds of instruments is determined by their respective uses. On occasion, one instrument may serve both purposes, acting as an aptitude test to forecast future performance and an achievement test to monitor past learning.

Creativity tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss76) assess a subject's ability to produce new ideas, insights, or artistic creations that are accepted as being of social, aesthetic, or scientific value. Thus, measures of **creativity** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss76) emphasize novelty and originality in the solution of fuzzy problems or the

production of artistic works. A creative response to one problem is illustrated in **Figure 1.1** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec4#ch01fig1)** .

Tests of creativity have a checkered history. In the 1960s, they were touted as a useful alternative to intelligence tests and used widely in U.S. school systems. Educators were especially impressed that creativity tests required divergent thinking—putting forth a variety of answers to a complex or fuzzy problem—as opposed to convergent thinking—finding the single correct solution to a well-defined problem. For example, a creativity test might ask the examinee to imagine all the things that would happen if clouds had strings trailing from them down to the ground. Students who could come up with a large number of consequences were assumed to be more creative than their less-imaginative colleagues. However, some psychometricians are skeptical, concluding that creativity is just another label for applied intelligence.

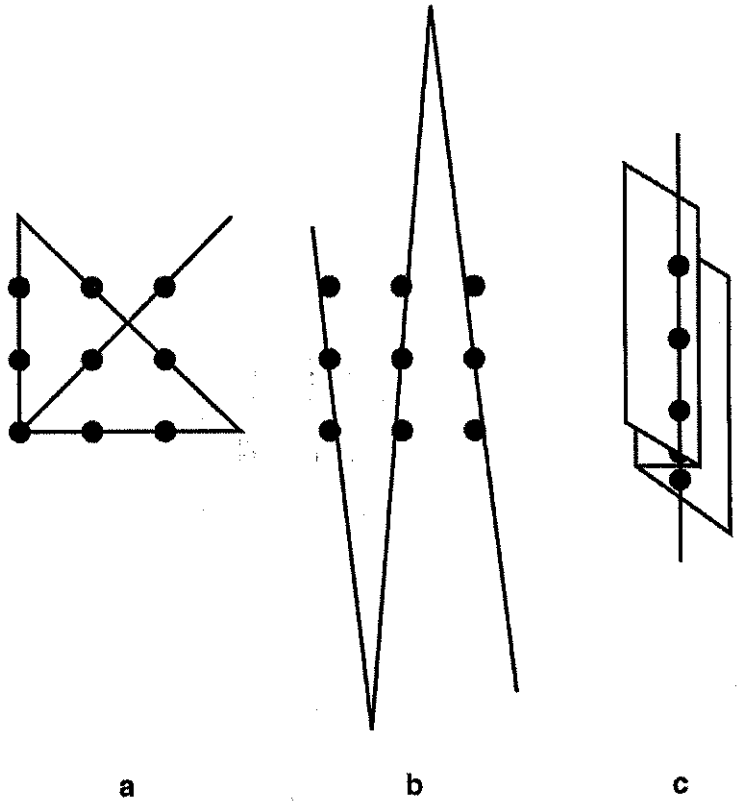


FIGURE 1.1 Solutions to the Nine-Dot Problem as Examples of Creativity

Note: Without lifting the pencil, draw through all the dots with as few straight lines as possible. The usual solution is shown in *a*. Creative solutions are depicted in *b* and *c*.

Personality tests **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss241)** measure the traits, qualities, or behaviors that determine a person's individuality; this information helps predict future behavior. These tests come in several different varieties, including checklists, inventories, and projective techniques such as sentence completions and inkblots **(Table 1.2** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec4#ch01tab2)** **)**.

Interest inventories **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss163)** measure an individual's preference for certain activities or topics and thereby help determine occupational choice. These tests are based on the explicit assumption that interest patterns determine and, therefore, also predict job satisfaction. For example, if the examinee has the same interests as successful and satisfied accountants, it is thought likely that he or she would enjoy the work of an accountant. The assumption that interest patterns predict job satisfaction is largely borne out by empirical studies, as we will review in a later chapter.

TABLE 1.2 Examples of Personality Test Items

(a) An Adjective Checklist	
Check those words which describe you:	
( ) relaxed	( ) assertive
( ) thoughtful	( ) curious
( ) cheerful	( ) even-tempered
( ) impatient	( ) skeptical
( ) morose	( ) impulsive
( ) optimistic	( ) anxious
(b) A True-False Inventory	

Circle true or false as each statement applies to you:

T	F	I like sports magazines.
T	F	Most people would lie to get a job.
T	F	I like big parties where there is lots of noisy fun.
T	F	Strange thoughts possess me for hours at a time.
T	F	I often regret the missed opportunities in my life.
T	F	Sometimes I feel anxious for no reason at all.
T	F	I like everyone I have met.
T	F	Falling asleep is seldom a problem for me.

(c) A Sentence Completion Projective Test

Complete each sentence with the first thought that comes to you:

I feel bored when

What I need most is

I like people who

My mother was

Many kinds of **behavioral procedures** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss31) are available for assessing the antecedents and consequences of behavior, including checklists, rating scales, interviews, and structured observations. These methods share a common assumption that behavior is best understood in terms of clearly defined characteristics such as frequency, duration, antecedents, and consequences. Behavioral procedures tend to be highly pragmatic in that they are usually interwoven with treatment approaches.

Neuropsychological tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss214) are used in the assessment of persons with known or suspected brain dysfunction. *Neuropsychology* is the study of brain–behavior relationships. Over the years, neuropsychologists have discovered that certain tests and procedures are highly sensitive to the effects of brain damage. Neuropsychologists use these specialized tests and procedures to make inferences about the locus, extent, and consequences of brain damage. A full neuropsychological assessment typically requires three to eight hours of one-on-one testing with an extensive battery of measures. Examiners must undergo comprehensive advanced training in order to make sense out of the resulting mass of test data.

1.5 USES OF TESTING

By far the most common use of psychological tests is to make decisions about persons. For example, educational institutions frequently use tests to determine placement levels for students, and universities ascertain who should be admitted, in part, on the basis of test scores. State, federal, and local civil service systems also rely heavily on tests for purposes of personnel selection.

Even the individual practitioner exploits tests, in the main, for decision making. Examples include the consulting psychologist who uses a personality test to determine that a police department hire one candidate and not another, and the neuropsycholo-gist who employs tests to conclude that a client has suffered brain damage.

But simple decision making is not the only function of psychological testing. It is convenient to distinguish five uses of tests:

- Classification
- Diagnosis and treatment planning
- Self-knowledge
- Program evaluation
- Research

These applications frequently overlap and, on occasion, are difficult to distinguish one from another. For example, a test that helps determine a psychiatric diagnosis might also provide a form of self-knowledge. Let us examine these applications in more detail.

The term **classification** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss47) encompasses a variety of procedures that share a common purpose: assigning a person to one category rather than another. Of course, the assignment to categories is not an end in itself but the basis for differential treatment of some kind. Thus, classification can have important effects such as granting or restricting access to a specific college or determining whether a person is hired for a particular job. There are many variant forms of classification, each emphasizing a particular purpose in assigning persons to categories. We will distinguish placement, screening, certification, and selection.

Placement (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss246) is the sorting of persons into different programs appropriate to their needs or skills. For example, universities often use a mathematics placement exam to determine whether students should enroll in calculus, algebra, or remedial courses.

Screening (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss287) refers to quick and simple tests or procedures to identify persons who might have special characteristics or needs. Ordinarily, psychometricians acknowledge that screening tests will result in many misclassifications. Examiners are, therefore, advised to do follow-up testing with additional instruments before making important decisions on the basis of screening tests. For example, to identify children with highly exceptional talent in spatial thinking, a psychologist might administer a 10-minute paper-and-pencil test to every child in a school system. Students who scored in the top 10 percent might then be singled out for more comprehensive testing.

Certification (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss45) and selection both have a pass/fail quality. Passing a certification exam confers privileges. Examples include the right to practice psychology or to drive a car. Thus, certification typically implies that a person has at least a minimum proficiency in some discipline or activity. Selection is similar to certification in that it confers privileges such as the opportunity to attend a university or to gain employment.

Another use of psychological tests is for diagnosis and treatment planning. **Diagnosis** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss91) consists of two intertwined tasks: determining the nature and source of a person's abnormal behavior, and classifying the behavior pattern within an accepted diagnostic system. Diagnosis is usually a precursor to remediation or treatment of personal distress or impaired performance.

Psychological tests often play an important role in diagnosis and treatment planning. For example, intelligence tests are absolutely essential in the diagnosis of mental retardation. Personality tests are helpful in diagnosing the nature and extent of emotional disturbance. In fact, some tests such as the MMPI were devised for the explicit purpose of increasing the efficiency of psychiatric diagnosis.

Diagnosis should be more than mere classification, more than the assignment of a label. A proper diagnosis conveys information—about strengths, weaknesses, etiology, and best choices for remediation/treatment. Knowing that a child has received a diagnosis of **learning disability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss183) is largely useless. But knowing in addition that the same child is well below average in reading comprehension, is highly distractible, and needs help with basic phonics can provide an indispensable basis for treatment planning.

Psychological tests also can supply a potent source of self-knowledge. In some cases, the feedback a person receives from psychological tests can change a career path or otherwise alter a person's life course. Of course, not every instance of psychological testing provides self-knowledge. Perhaps in the majority of cases the client already knows what the test results divulge. A high-functioning college student is seldom surprised to find that his IQ is in the superior range. An architect is not perplexed to hear that she has excellent spatial reasoning skills. A student with meager reading capacity is usually not startled to receive a diagnosis of "learning disability."

Another use for psychological tests is the systematic evaluation of educational and social programs. We have more to say about the evaluation of educational programs when we discuss achievement tests in a later chapter. We focus here on the use of tests in the evaluation of social programs. Social programs are designed to provide services that improve social conditions and community life. For example, Project Head Start is a federally funded program that supports nationwide pre-school teaching projects for underprivileged children (**McKey and others, 1985** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1104)). Launched in 1965 as a precedent-setting attempt to provide child development programs to low-income families, Head Start has provided educational enrichment and health services to millions of at-risk preschool children.

But exactly what impact does the multi-billion-dollar Head Start program have on early childhood development? Congress wanted to know if the program improved scholastic performance and reduced school failure among the enrollees. But the centers vary by sponsoring agencies, staff characteristics, coverage, content, and objectives, so the effects of Head Start are not easy to ascertain. Psychological tests provide an objective basis for answering these questions that is far superior to anecdotal or impressionistic reporting. In general, Head Start children show immediate gains in IQ, school readiness, and academic achievement, but these gains dissipate in the ensuing years (**Figure 1.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec5#ch01fig2)).

So far we have discussed the practical application of psychological tests to everyday problems such as job selection, diagnosis, or program evaluation. In each of these instances, testing serves an immediate, pragmatic purpose: helping the tester make decisions about persons or programs. But tests also play a major role in both the applied and theoretical branches of behavioral research. As an example of testing in applied research, consider the problem faced by neuropsychologists who wish to investigate the hypothesis that low-level lead absorption causes behavioral deficits in children. The only feasible way to explore this supposition is by testing normal and lead-burdened children with a battery of psychological tests. Needleman and associates (**1979** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1220**)) used an array of traditional and innovative tests to conclude that low-level lead absorption causes decrements in IQ, impairments in reaction time, and escalations of undesirable classroom behaviors. Their conclusions inspired a tumultuous and bitter exchange of opinions that we will not review here (**Needleman et al., 1990** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1221**)). However, the passions inspired by this study epitomize an instructive point: Academicians and public policymakers respect psychological tests. Why else would they engage in lengthy, acrimonious debates about the validity of testing-based research findings?

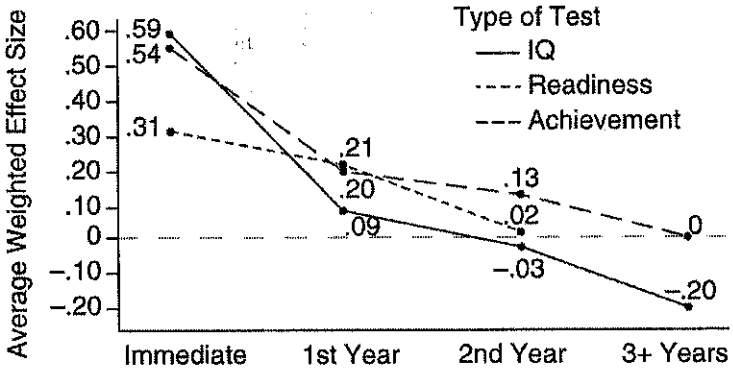


FIGURE 1.2 Longitudinal Test Results from the Head Start Project

Source: From McKey, R. H., and others. (1985). *The Impact of Head Start on children, families and communities.* Washington, DC: U.S. Government Printing Office. In the public domain.

1.6 FACTORS INFLUENCING THE SOUNDNESS OF TESTING

Psychological testing is a dynamic process influenced by many factors. Although examiners strive to ensure that test results accurately reflect the traits or capacities being assessed, many extraneous factors can sway the outcome of psychological testing. In this section, we review the potentially crucial impact of several sources of influence: the manner of administration, the characteristics of the tester, the context of the testing, the motivation and experience of the examinee, and the method of scoring.

The sensitivity of the testing process to extraneous influences is obvious in cases where the examiner is cold, hurried, or incompetent. However, invalid test results do not originate only from obvious sources such as blatantly nonstandard administration, hostile tester, noisy testing room, or fearful examinee. In addition, there are numerous, subtle ways in which method, examiner, context, or motivation can alter test results. We provide a comprehensive survey of these extraneous influences in the remainder of this topic.

1.7 STANDARDIZED PROCEDURES IN TEST ADMINISTRATION

The interpretation of a psychological test is most reliable when the measurements are obtained under the standardized conditions outlined in the publisher's test manual. Nonstandard testing procedures can alter the meaning of the test results, rendering them invalid and, therefore, misleading. Standardized procedures are so important that they are listed as an essential criterion for valid testing in the *Standards for Educational and Psychological Testing* (**1999 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30)**), a reference manual published jointly by the American Psychological Association and other groups:

> In typical applications, test administrators should follow carefully the standardized procedures for administration and scoring specified by the test publisher. Specifications regarding instructions to test takers, time limits, the form of item presentation or response, and test materials or equipment should be strictly observed. Exceptions should be made only on the basis of carefully considered professional judgment, primarily in clinical applications. **(AERA, APA, NCME, 1999 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30))**

Suppose the instructions to the vocabulary section of a children's intelligence test specify that the examiner should ask, "What does *sofa* mean, what is a sofa?" If a subject were to reply, "I've never heard that word," an inexperienced tester might be tempted to respond, "You know, a couch—what is a couch?" This may strike the reader as a harmless form of fair play, a simple rephrasing of the original question. Yet, by straying from standardized procedures, the examiner has really given a different test. The point in asking for a definition of *sofa* (and not *couch*) is precisely that *sofa* is harder to define and, therefore, a better index of high-level vocabulary skills.

Even though standardized testing procedures are normally essential, there are instances in which flexibility in procedures is desirable or even necessary. As suggested in the APA *Standards*, such deviations should be reasoned and deliberate. An analogy to the spirit of the law versus the letter of the law is relevant here. An overly zealous examiner might capture the letter of the law, so to speak, by adhering literally and strictly to testing procedures outlined in the publisher's manual. But is this really what most test publishers intend? Is it even how the test was actually administered to the normative sample? Most likely publishers would prefer that examiners capture the spirit of the law even if, on occasion, it is necessary to adjust testing procedures slightly.

The need to adjust standardized procedures for testing is especially apparent when examining persons with certain kinds of disabilities. A subject with a speech impediment might be allowed to write down the answers to orally presented questions or to use gesture and pantomime in response to some items. For example, a test question might ask, "What shape is a ball?" The question is designed to probe the subject's knowledge of common shapes, not to examine whether the examinee can verbalize "round." The written response *round* and the gestured response (a circular motion of the index finger) are equally correct, too.

Minor adjustments in procedures that heed the spirit in which a test was developed occur on a regular basis and are no cause for alarm. These minor adjustments do not invalidate the established norms—on the contrary, the appropriate adaptation of procedures is necessary so that the norms remain valid. After all, the testers who collected data from the standardization sample did not act like heartless robots when posing questions to subjects. Examiners who wish to obtain valid results must likewise exercise a reasoned flexibility in testing procedures.

However, considerable clinical experience is needed to determine whether an adjustment in procedure is minor or so substantial that existing norms no longer apply. This is why psychological examiners normally receive extensive supervised experience before they are allowed to administer and interpret individual tests of ability or personality.

In certain cases an examiner will knowingly depart from standard procedures to a substantial degree; this practice precludes the use of available test norms. In these instances, the test is used to help formulate clinical judgments rather than to determine a quantitative index. For example, when examining aphasic patients, it may be desirable to ignore time limits entirely and accept roundabout answers. The examiner might not even calculate a score. In these rare cases, the test becomes, in effect, an adjunct to the clinical interview. Of course, when the examiner does not adhere to standardized procedures, this should be stated explicitly in the written report.

1.8 DESIRABLE PROCEDURES OF TEST ADMINISTRATION

A small treatise could be written on desirable procedures of test administration, but we will have to settle for a brief listing of the most essential points. For more details, the interested reader can consult Sattler (**2001** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1438)) on the individual testing of children and Clemans (**1971** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib314)) on group testing. We discuss individual testing first, then briefly list some important points about desirable procedures in group testing.

An essential component of individual testing is that examiners must be intimately familiar with the materials and directions before administration begins. Largely this involves extensive rehearsal and anticipation of unusual circumstances and the appropriate response. A well-prepared examiner has memorized key elements of verbal instructions and is ready to handle the unexpected.

The uninitiated student of assessment often assumes that examination procedures are so simple and straightforward that a quick once-through reading of the manual will suffice as preparation for testing. Although some individual tests are exceedingly rudimentary and uncomplicated, many of them have complexities of administration that, unheeded, can cause the examinee to fail items unnecessarily. For example, Choi and Proctor (**1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib304)) found that 25 of 27 graduate students made serious errors in the administration of the Stanford-Binet: Fourth Edition, even though the sessions were videotaped and the students knew their testing skills were being evaluated. Ramos, Alfonso, and Schermerhorn (**2009** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1331)) reviewed 108 protocols from the Woodcock Johnson III Tests of Cognitive Abilities administered by 36 first-year graduate students in a school psychology doctoral program. The researchers found an average of almost 5 errors per test, including the use of incorrect ceilings, failure to record errors, and failure to encircle the correct row for the total number correct. Loe, Kadlubek, and Williams (**2007** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1001)) reviewed 51 WISC-IV protocols administered by graduate students and found an average of almost 26 errors per protocol. The two most common errors were the failure to query incomplete or ambiguous verbal responses, and granting too many points for substandard answers. In many cases, these errors materially affected the Full Scale IQ, shifting it upward or downward from the likely true score. What these studies confirm is that appropriate attention to the details of administration and scoring is essential for valid results.

The necessity for intimate familiarity with testing procedures is well illustrated by the Block Design subtest of the WAIS-IV (**Wechsler, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1737)). The materials for the subtest include nine blocks (cubes) colored red on two sides, white on two sides, and red/white on two sides. The examinee's task is to use the blocks to construct patterns depicted on cards. For the initial designs, four blocks are needed, while for more difficult designs, all nine blocks are provided (**Figure 1.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec8#ch01fig3)).

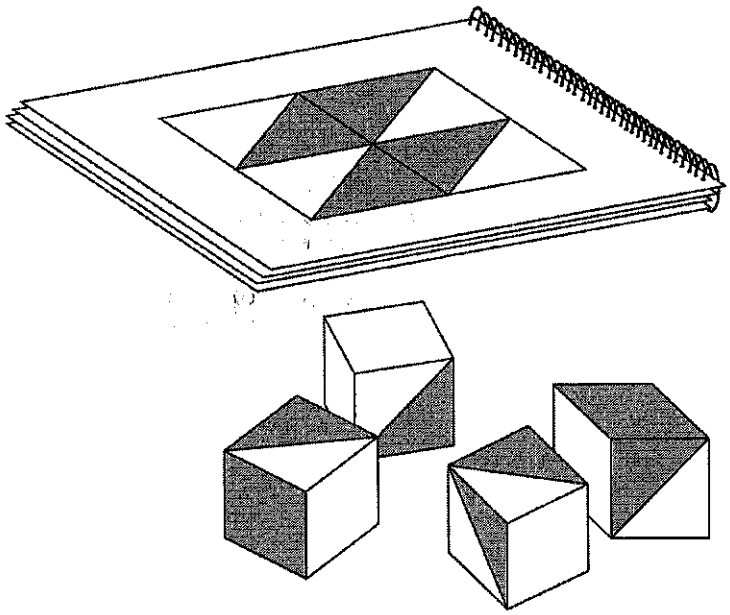


FIGURE 1.3 Materials Similar to WAIS-IV Block Design Subtest

Bright examinees have no difficulty comprehending this task and the exact instructions do not influence their performance appreciably. However, persons whose intelligence is average or below average need the elaborate demonstrations and corrections that are specified in the WAIS-IV manual (**Wechsler, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1737)). In particular, the examiner demonstrates the first two designs and responds to the examinee's success or failure on these according to a complex flow of reaction and counterreaction, as outlined in *three pages* of instructions. Woe to the tester who has not rehearsed this subtest and anticipated the proper response to examinees who falter on the first two designs.

Sensitivity to Disabilities

Another important ingredient of valid test administration is sensitivity to disabilities in the examinee. Impairments in hearing, vision, speech, or motor control may seriously distort test results. If the examiner does not recognize the physical disability responsible for the poor test performance, a subject may be branded as intellectually or emotionally impaired when, in fact, the essential problem is a sensory or motor disability.

Vernon and Brown (**1964** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1694)) reported the tragic case of a young girl who was relegated to a hospital for the mentally retarded as a consequence of the tester's insensitivity to physical disability. The examiner failed to notice that the child was deaf and concluded that her Stanford-Binet IQ of 29 was valid. She remained in the hospital for five years, but was released after she scored an IQ of 113 on a performance-based intelligence test! After dismissal from the hospital, she entered a school for the deaf and made good progress.

Persons with disabilities may require specialized tests for valid assessment. The reader will encounter a lengthy discussion of available tests for exceptional examinees in **Chapter 7** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch07#ch07) , Testing Special Populations. In this section, we concentrate on the vexing issues raised when standardized tests for normal populations are used with mildly or moderately disabled subjects. We include separate discussions of the testing process for examinees with a hearing, vision, speech, or motor control problem. However, the reader needs to know that many exceptional examinees have multiple disabilities.

Valid testing of a subject with a hearing impairment requires first of all that the examiner detect the existence of the disability! This is often more difficult than it seems. Many persons with mild hearing loss learn to compensate for this disability by pretending to understand what others say and waiting for further conversational cues to help clarify faintly perceived words or phrases. As a result, other persons—including psychologists—may not perceive that an individual with mild hearing loss has any disability at all.

Failure to notice a hearing loss is particularly a problem with young examinees, who are usually poor informants about their disabilities. Young children are also prone to fluctuating hearing losses due to the periodic accumulation of fluid in the middle ear during intervals of mild illness (**Vernon & Alles, 1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1693)). A child with a fluctuating hearing loss may have normal hearing in the morning, but perceive conversational speech as a whisper just a few hours later.

Indications of possible hearing difficulty include lack of normal response to sound, inattentiveness, difficulty in following oral instructions, intent observation of the speaker's lips, and poor articulation (**Sattler, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1437)). In all cases in which hearing impairment is suspected, referral for an audiological examination is crucial. If a serious hearing problem is confirmed, then the examiner should consider using one of the specialized tests discussed in **Chapter 7** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch07#ch07) , Testing Special Populations. In persons with a mild hearing loss, it is essential for the examiner to face the subject squarely, speak loudly, and repeat instructions slowly. It is also important to find a quiet room for testing. Ideally, a testing room will have curtains and textured wall surfaces to minimize the distracting effects of background noises.

In contrast to those with hearing loss, subjects with visual disabilities generally attend well to verbally presented test materials. The examinee with visual impairment introduces a different kind of challenge to the examiner: detecting that a visual impairment exists, and then ensuring that the subject can see the test materials well.

Detecting visual impairment is a straightforward matter with adult subjects—in most cases, a mature examinee will freely volunteer information about visual impairment, especially if asked. However, children are poor informants about their visual capacities, so testers need to know the signs and symptoms of possible visual impairment in a young examinee. Common sense is a good starting point: Children who squint, blink excessively, or lose their place when reading may have a vision problem. Holding books or testing materials up close is another suspicious sign. Blurred or double vision may signify visual problems, as may headaches or nausea after reading. In general, it is so common for children to require corrective lenses that examiners should be on the lookout for a vision problem in any young subject who does not wear glasses and has not had a recent vision exam.

Depending on the degree of visual impairment, examiners need to make corresponding adjustments in testing. If the child's vision is of no practical use, special instruments with appropriate norms must be used. For example, the Perkins-Binet is available for testing children who are blind. These tests are discussed in **Topic 7B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch07lev1sec5#ch07box2) , Testing Persons with Disabilities. For obvious reasons, only the verbal portions of tests should be administered to sighted children with an uncorrected visual problem.

Speech impairments present another problem for diagnosticians. The verbal responses of subjects with speech impairment are difficult to decipher. Owing to the failed comprehension of the examiner, subjects may receive less credit than is due. Sattler (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1437)) relates the lamentable case of Daniel Hoffman, a youngster with speech impairment who spent his entire youth in classes for those with mental retardation because his Stanford-Binet IQ was 74. In actuality, his intelligence was within the normal range, as revealed by other performance-based tests. In another tragic miscarriage of assessment, a patient in England was mistakenly confined to a ward for those with severe retardation because cerebral palsy rendered his speech incomprehensible. The patient was wheelchair-bound and had almost no motor control, so his performance on nonverbal tests was also grossly impaired. The staff assumed he was severely retarded, so the patient remained on the back ward for decades. However, he befriended a fellow resident who could comprehend the patient's gutteral rendition of the alphabet. The friend was severely retarded but could nonetheless recognize keys on a typewriter. With laborious letter-by-letter effort, the patient with incapacitating cerebral palsy wrote and published an autobiography, using his friend with mental disability as a conduit to the real world.

Even if their disability is mild, persons with cerebral palsy or other motor impairments may be penalized by timed performance tests. When testing a person with a mild motor disability, examiners may wish to omit timed performance subtests or to discount these results if they are consistently lower than scores from untimed subtests. If a subject has an obvious motor disability—such as a difficulty in manipulating the pieces of a puzzle—then standard instruments administered in the normal manner are largely inappropriate. A number of alternative instruments have been developed expressly for examinees with cerebral palsy and other motor impairments, and standard tests have been cleverly adapted and renormed (**Topic 7B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch07lev1sec5#ch07box2) , Testing Persons with Disabilities).

Desirable Procedures of Group Testing

Psychologists and educators commonly assume that almost any adult can accurately administer group tests, so long as he or she has the requisite manual. Administering a group test would appear to be a simple and straightforward procedure of passing out forms and pencils, reading instructions, keeping time, and collecting the materials.

In reality, conducting a group test requires as much finesse as administering an individual test, a point recognized years ago by Traxler (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1662)). There are numerous ways in which careless administration and scoring can impair group test results, causing bias for the entire group or affecting only certain individuals. We outline only the more important inadequacies and errors in the following paragraphs, referring the reader to Traxler (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1662)) and Clemans (**1971** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib314)) for a more complete discussion.

Undoubtedly the greatest single source of error in group test administration is incorrect timing of tests that require a time limit. Examiners must allot sufficient time for the entire testing process: setup, reading instructions out loud, and the actual test taking by examinees. Allotting sufficient time requires foresightful scheduling. For example, in many school settings, children must proceed to the next class at a designated time, regardless of ongoing activities. Inexperienced examiners might be tempted to cut short the designated time limit for a test so that the school schedule can be maintained. Of course, reduced time on a test renders the norms completely invalid and likely lowers the score for most subjects in the group.

Allowing too much time for a test can be an equally egregious error. For example, consider the impact of receiving extra time on the Miller Analogies Test (MAT), a high-level reasoning test once required by many universities for graduate school application. Since the MAT is a speeded test that requires quick analogical thinking, extra time would allow most examinees to solve several extra problems. This kind of testing error would likely lower the validity of the MAT results as a predictor of graduate school performance.

A second source of error in group test administration is lack of clarity in the directions to the examinees. Examiners must read the instructions slowly in a clear, loud voice that commands the attention of the subjects. Instructions must not be paraphrased. Where allowed by the manual, examiners must stop and clarify points with individual examinees who are confused.

Noise is another factor that must be controlled in group testing. It has been known for some time that noise causes a decrease in performance, especially for tasks of high complexity (**e.g., Boggs & Simon, 1968** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib174)). Surprisingly, there is little research on the effects of noise on psychological tests. However, it seems almost certain that loud noise, especially if intermittent and unpredictable, will cause test scores to decline substantially. Elementary schoolchildren should not be expected to perform well while a construction worker jackhammers a cement wall in the next room. In fairness to the examinees, there are times when the test administrator should reschedule the test.

Another source of error in the administration of a group test is failure to explain when and if examinees should guess. Perhaps more frequently than any other question, examiners are asked, "Is there a penalty if I guess wrong?" In most instances, test developers anticipate this issue and provide explicit guidance to subjects as to the advantages and/or pitfalls of guessing. Examiners should not give supplementary advice on guessing—this would constitute a serious deviation from standardized procedure.

Most test developers incorporate a **correction for guessing** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss72) based on established principles of probability. Consider a multiple-choice test that has four alternatives per item. On those items that the subject makes a wild, uneducated guess, the odds on being correct are 1 out of 4, while the odds on being wrong are 3 out of 4. Thus, for every three wrong guesses, there will be one correct guess that reflects luck rather than knowledge. Suppose a young girl answers correctly on 35 questions from a 50-item test but answers erroneously on 9 questions. In all, she has answered 44 questions, leaving 6 blank. The fact that she selected the wrong alternative on 9 questions suggests that she also gained 3 correct answers due to luck rather than knowledge. Remember, on wild guesses we expect there to be, on average, 3 wrong answers for every correct answer, so for 9 wrong guesses we would expect 3 correct guesses on other questions. The subject's corrected score—the one actually reported and compared to existing norms—would then be 32; that is, 35 minus 3. In other words, she probably knew 32 answers but by guessing on 12 others she boosted her score another 3 points.

The scoring correction outlined in the preceding paragraph pertains only to wild, uneducated guesses. The effect of such a correction is to eliminate the advantage otherwise bestowed on unabashed risk takers. However, not all guesses are wild and uneducated. In some instances, an examinee can eliminate one or two of the alternatives, thereby increasing the odds of a correct guess among the remaining choices. In this situation, it may be wise for the examinee to guess.

Whether an educated guess is really to the advantage of the examinee depends partly on the diabolical skill of the item writer. Traxler (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1662)) notes:

> In effect, the item writer attempts to make each wrong response so plausible that every examinee who does not possess the desired skill or ability will select a wrong response. In other words, the item writer's aim is to make all or nearly all considered guesses wrong guesses.

A skilled item writer can fashion questions so that the correct alternative is completely counterintuitive and the wrong alternatives are persuasively appealing. For these items, an educated guess is almost always wrong.

Nonetheless, many test developers now advise subjects to make educated guesses but warn against wild guesses. For example, a recent edition of the test preparation manual *Taking the SAT* advises:

> Because of the way the test is scored, haphazard or random guessing for questions you know nothing about is unlikely to change your score. When you know that one or more choices can be eliminated, guessing from among the remaining choices should be to your advantage.

Whether or not a group test uses a scoring correction, the important point to emphasize in this context is that the administrator should follow standardized procedure and never offer supplementary advice about guessing. In group testing, deviations from the instructions manual are simply unacceptable.

1.9 INFLUENCE OF THE EXAMINER

The Importance of Rapport

Test publishers urge examiners to establish **rapport** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss266) —a comfortable, warm atmosphere that serves to motivate examinees and elicit cooperation. Initiating a cordial testing milieu is a crucial aspect of valid testing. A tester who fails to establish rapport may cause a subject to react with anxiety, passive-aggressive noncooperation, or open hostility. Failure to establish rapport distorts test findings: Ability is underestimated and personality is misjudged.

Rapport is especially important in individual testing and particularly so when evaluating children. Wechsler (**1974** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1731)) has noted that establishing rapport places great demands on the clinical skills of the tester:

> To put the child at ease in his surroundings, the examiner might engage him in some informal conversation before getting down to the more serious business of giving the test. Talking to him about his hobbies or interests is often a good way of breaking the ice, although it may be better to encourage a shy child to talk about something concrete in the environment—a picture on the wall, an animal in his classroom, or a book or toy (not a test material) in the examining room. In general, this introductory period need not take more than 5 to 10 minutes, although the testing should not start until the child seems relaxed enough to give his maximum effort.

Testers may differ in their abilities to establish rapport. Cold testers will likely obtain less cooperation from their subjects, resulting in reduced performance on ability tests or distorted, defensive results on personality tests. Overly solicitous testers may err in the opposite direction, giving subtle (and occasionally blatant) cues to correct answers. Both extremes should be avoided.

Examiner Sex, Experience, and Race

A wide body of research has sought to determine whether certain characteristics of the examiner cause examinee scores to be raised or lowered on ability tests. For example, does it matter whether the examiner is male or female? Experienced or novice? Same or different race from the examinee? We will contain the urge to review these studies—with a few exceptions—for one simple reason: The results are contradictory and, therefore, inconclusive. Most studies find that sex, experience, and race of the examiner make little, if any, difference. Furthermore, the few studies that report a large effect in one direction (e.g., female examiners elicit higher IQ scores) are contradicted by other studies showing the opposite trend. The interested reader can consult Sattler (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1437)) for a discussion and extensive listing of references.

Yet, it would be unwise to conclude that sex, experience, or race of the examiner never affect test scores. In isolated instances, a particular examiner characteristic might very well have a large effect on examinee test scores. For example, Terrell, Terrell, and Taylor (**1981** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1631)) ingeniously demonstrated that the race of the examiner interacts potently with the trust level of African American examinees in IQ testing. These researchers identified African American college students with high and low levels of mistrust of whites; half of each group was then administered the WAIS by a white examiner, the other half by an African American examiner. The high-mistrust group with an African American examiner scored significantly higher than the high-mistrust group with a white examiner (average IQs of 96 versus 86, respectively). In addition, the low-mistrust group with a white examiner scored slightly higher than the low-mistrust group with an African American examiner (average IQs of 97 versus 92, respectively). In sum, the authors concluded that mistrustful African Americans do poorly when tested by white examiners. Data bearing on this type of racial effect are meager, and there is certainly room for additional research.

1.10 BACKGROUND AND MOTIVATION OF THE EXAMINEE

Examinees differ not only in the characteristics that examiners desire to assess but also in other extraneous ways that might confound the test results. For example, a bright subject might perform poorly on a speeded ability test because of test anxiety; a sane murderer might seek to appear mentally ill on a personality inventory to avoid prosecution; a student of average ability might undergo coaching to perform better on an aptitude test. Some subjects utterly lack motivation and don't care if they do well on psychological tests. In all of these instances, the test results may be inaccurate because of the filtering and distorting effects of certain examinee characteristics such as anxiety, malingering, coaching, or cultural background.

Test Anxiety

Test anxiety (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss327) refers to those phenomenological, physiological, and behavioral responses that accompany concern about possible failure on a test. There is no doubt that subjects experience different levels of test anxiety ranging from a carefree outlook to incapacitating dread at the prospect of being tested.

Several true-false questionnaires have been developed to assess individual differences in test anxiety (e.g., **Lowe, Lee, Witteborg, & others, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1010) ; **Spielberger, Gonzalez, Taylor, & others, 1980** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1536) ; **Spielberger & Vagg, 1995** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1535)). Following, we list characteristic items and their direction of keying (*T* for True, *F* for False):

- (T) When taking an important examination, I sweat a great deal.
- (T) I freeze up when I take intelligence tests or school exams.
- (F) I really don't understand why some people get so upset about tests.
- (T) I dread courses in which the instructor likes to give "pop" quizzes.

An extensive body of research has confirmed the commonsense notion that test anxiety is negatively correlated with school achievement, aptitude test scores, and measures of intelligence (e.g., **Chapell, Blanding, & Silverstein, 2005** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib297) ; **Naveh-Benjamin, McKeachie, & Lin, 1987** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1219) ; **Ortner & Caspers, 2011** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1258)). However, the interpretation of these correlational findings is not straightforward. One possibility is that students develop test anxiety because of a history of performing poorly on tests. That is, the decrements in performance may precede and cause the test anxiety. In support of this viewpoint, Paulman and Kennelly (**1984** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1274)) found that—independent of their anxiety—many test-anxious students also display ineffective test taking in academic settings. Such students would do poorly on tests whether or not they were anxious. Moreover, Naveh-Benjamin et al. (**1987** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1219)) determined that a large proportion of test-anxious college students have poor study habits that predispose them to poor test performance. The test anxiety of these subjects is partly a by-product of lifelong frustration over mediocre test results.

Other lines of research indicate that test anxiety has a directly detrimental effect on test performance. That is, test anxiety is likely both cause and effect in the equation linking it with poor test performance. Consider the seminal study on this topic by Sarason (**1961** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1435)), who tested high- and low-anxious subjects under neutral or anxiety-inducing instructions. The subjects were college students required to memorize two-syllable words low in meaningfulness—a difficult task. Half of the subjects performed under neutral instructions—they were simply told to memorize the lists. The remaining subjects were told to memorize the lists and told that the task was an intelligence test. They were urged to perform as well as possible. The two groups did not differ significantly in performance when the instructions were neutral and non-threatening. However, when the instructions aroused anxiety, performance levels for the high-anxious subjects dropped markedly, leaving them at a huge disadvantage compared to low-anxious subjects. This indicates that test-anxious subjects show significant decrements in performance when they perceive the situation as a test. In contrast, low-anxious subjects are relatively unaffected by such a simple redefinition of the context.

Tests with narrow time limits pose a special problem to persons with high levels of test anxiety. Time pressure seems to exacerbate the degree of personal threat, causing significant reductions in the performance of test-anxious persons. Siegman (**1956** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1497)) demonstrated this point many years ago by comparing performance levels of high- and low-anxious medical/psychiatric patients on timed and untimed subtests from the WAIS. The WAIS consists of eleven subtests, including six subtests for which the examiner uses a stopwatch to enforce strict time limits, and five subtests for which the subject has unlimited time to respond. Interestingly, the high- and low-anxious subjects were of equal overall ability on the WAIS. However, each group excelled on different kinds of subtests in predictable directions. In particular, the low-anxious subjects surpassed the high-anxious subjects on timed subtests, whereas the reverse pattern was observed on untimed subtests (**Figure 1.4** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec10#ch01fig4)).

Motivation to Deceive

Test results also may be inaccurate if the examinee has reasons to perform in an inadequate or unrepresentative manner. Overt faking of test results is rare, but it does happen. A small fraction of persons seeking benefits from rehabilitation or social agencies will consciously fake bad on personality and ability tests. The topic of malingering (faking bad for personal gain) is discussed in a later chapter.

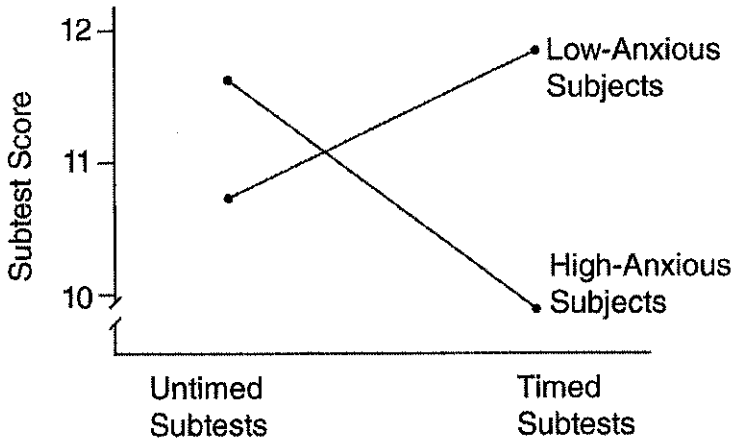


FIGURE 1.4 Influence of Timing and Anxiety Level on WAIS Subtest Results

Source: Based on data from Siegman, A. W. (1956). The effect of manifest anxiety on a concept formation task, a nondirected learning task, and on timed and untimed intelligence tests. *Journal of Consulting Psychology, 20,* 176–178.

TOPIC 1B Ethical and Social Implications of Testing

1.11 The Rationale for Professional Testing Standards (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec11#ch01lev1sec11)

Case Exhibit 1.2 Ethical and Professional Quandaries in Testing
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec11#ch01exh2)

1.12 Responsibilities of Test Publishers (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec12#ch01lev1sec12)

1.13 Responsibilities of Test Users (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec13#ch01lev1sec13)

Case Exhibit 1.3 Overzealous Interpretation of the MMPI (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec13#ch01exh3)

1.14 Testing of Cultural and Linguistic Minorities (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec14#ch01lev1sec14)

1.15 Unintended Effects of High-Stakes Testing (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec15#ch01lev1sec15)

1.16 Reprise: Responsible Test Use (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec16#ch01lev1sec16)

The general theme of this book is that psychological testing is a beneficial influence in modern society. When used ethically and responsibly, testing provides a basis for arriving at sensible inferences about individuals and groups. After all, the intention of the enterprise is to promote proper guidance, effective treatment, accurate evaluation, and fair decision making—whether in one-on-one clinic testing or institutional group testing. Who could possibly complain about these goals?

Thankfully, tests generally are applied in an ethical and responsible manner by psychologists, educators, administrators, and others. But there are exceptions. Almost everyone has heard the horrific anecdotes: the minority grade schooler casually labeled as having mental retardation on the basis of a single IQ score; the college student implausibly diagnosed as schizophrenic from a projective test; the job applicant wrongfully screened from employment based on an irrelevant measure; the aspiring teacher given unfair advantage when a competency test is mysteriously leaked beforehand; or the minority child penalized in testing because English is not her first language. Exceptions such as these illustrate the need for ethical and professional standards in testing.

A major purpose of this topic is to introduce the reader to the ethical and professional standards that inform the practice of psychological testing. We also pursue the related theme of special considerations in the testing of cultural and linguistic minorities. The two topics share substantial overlap: When an examinee is not from the majority Anglo-American culture (predominantly Caucasian, English-speaking, individualistic, future-oriented), ethical and professional concerns in testing rise to the forefront.

Finally, we examine a troubling and under-reported implication of widespread testing, namely, to the extent that society uses test results to make important decisions, the motivation for stakeholders to cheat is intensified. As a result, cheating has emerged as a dark, unintended consequence of high-stakes testing, especially in the school systems of our nation.

1.11 THE RATIONALE FOR PROFESSIONAL TESTING STANDARDS

Testing is generally applied in a responsible manner, but as previously noted, there are exceptions. On rare occasions, testing is irresponsible by design rather than by accident. Consider, with shuddering amazement, the advertisement for Mind Prober featured in a pop psychology magazine:

> Read Any Good Minds Lately? With the Mind Prober you can. In just minutes you can have a scientifically accurate personality profile of anyone. This new expert systems software lets you discover the things most people are afraid to tell you. The strengths, weaknesses, sexual interests and more. **(Eyde & Primhoff, 1992 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib472))**

In this case the irresponsibility is so blatant that discussion of ethical and professional guidelines is almost superfluous.

However, testing practices do not always present in sharply contrasting shades, responsible or irresponsible. The real challenge of competent assessment is to determine the boundaries of ethical and professional practice. As usual, it is the borderline cases that provide pause for thought. The reader is encouraged to read the quandaries of testing described in **Case Exhibit 1.2 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec11#ch01exh2)** and form an opinion about each. These examples are based on firsthand reports to the author. At the close of this chapter, we will return to these problematic vignettes.

CASE EXHIBIT 1.2

Ethical and Professional Quandaries in Testing

- A consulting psychologist agrees to perform preemployment screening for psychopathology in police officer candidates. At the beginning of each consultation, the psychologist asks the candidate to read and sign a detailed consent form that openly and honestly describes the evaluation process. However, the consent form explains that specific feedback about the test results will not be provided to job candidates. Question: Is it ethical for the psychologist to deny such feedback to the candidates?
- A competent counselor who has received extensive training in the interpretation of the MMPI continues to use this instrument even though it has been superceded by the MMPI-2. His rationale is simply that there is a huge body of research on the MMPI and, he feels secure about the meaning of elevated MMPI test profiles, whereas he knows very little about the MMPI-2. He intends to switch over to the MMPI-2 at some undetermined future date, but finds no compelling reason to do so immediately. Question: Is the counselor's refusal to use the MMPI-2 a breach of professional standards?
- A consulting psychologist is asked to evaluate a 9-year-old boy of Puerto Rican descent for possible learning disability. The child's primary language is Spanish and his secondary language is English. The psychologist intends to use the Wechsler Intelligence Scale for Children-IV (WISC-IV) and other tests. Because he knows almost no Spanish, the psychologist asks the child's after-school babysitter to act as translator when this is required to communicate test directions, specific questions, or the child's responses. Question: Is it an appropriate practice to use a translator when administering an individual test such as the WISC-IV?
- In the midst of taking a test battery for learning disability, a distraught 20-year-old female college student confides a terrifying secret to the psychologist. The client has just discovered that her 25-year-old brother, who died three months ago, was most likely a pedophile. She shows the psychologist photographs of naked children posing in the brother's bedroom. To complicate matters, the brother lived with his mother—who is still unaware of his well-concealed sexual deviancy. Question: Is the psychologist obligated to report this case to law enforcement?

The dilemmas of psychological testing do not always have simple, obvious answers. Even thoughtful and experienced psychologists may disagree as to what is ethical or professional in a given instance. Nonetheless, the scope of ethical and professional practice is not a matter of individual taste or personal judgment. Responsible test use is defined by written guidelines published by professional associations such as the American Psychological Association, the American Counseling Association, the National Association of School Psychologists, and other groups. Whether they know it or not, all practitioners owe allegiance to these guidelines, which we review in the following sections.

In general, the evolution of professional and ethical standards has been almost uniformly restrictive, providing an ever-narrowing demarcation of where, when, and how psychological tests may be used. Partly in response to the modern climate of litigation, organizations concerned with psychological testing have published guidelines that collectively define the ethical and professional standards relevant to the practice of assessment.

These standards also pertain to corporations and individuals who publish tests. We begin with a survey of guidelines for test publishers before examining the responsibilities of test users. The chapter closes with a review of special concerns in the testing of cultural and linguistic minorities.

1.12 RESPONSIBILITIES OF TEST PUBLISHERS

The responsibilities of publishers pertain to the publication, marketing, and distribution of their tests. In particular, it is expected that publishers will release tests of high quality, market their product in a responsible manner, and restrict distribution of tests only to persons with proper qualifications. We consider each of these points in turn.

Publication and Marketing Issues

Regarding the publication of new or revised instruments, the most important guideline is to guard against premature release of a test. Testing is a noble enterprise but it is also big business driven by the profit motive, which provides an inherent pressure toward early release of new or revised materials. Perhaps this is why the American Psychological Association and other organizations have published standards that relate to test publication (**AERA/APA/NCME, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30)). These standards pertain especially to the technical manuals and user guides that typically accompany a test. These sources must be sufficiently complete so that a qualified user or reviewer can evaluate the appropriateness and technical adequacy of the test. This means that manuals and guides will report detailed statistics on reliability analyses, validity studies, normative samples, and other technical aspects.

Marketing tests in a responsible manner refers not only to advertising (which should be accurate and dignified) but also to the way in which information is portrayed in manuals and guides. In particular, test authors should strive for a balanced presentation of their instruments and refrain from a one-sided presentation of information. For example, if some preliminary studies reflect poorly on a test, these should be given fair weight in the manual alongside positive findings. Likewise, if a potential misuse or inappropriate use of a test can be anticipated, the test author needs to discuss this matter as well.

Competence of Test Purchasers

Test publishers recognize the broad responsibility that only qualified users should be able to purchase their products. By way of brief review, the reasons for restricted access include the potential for harm if tests fall into the wrong hands (e.g., an undergraduate psychology major administers the MMPI-2 to his friends and then makes frightful pronouncements about the results) and the obvious fact that many tests are no longer valid if potential examinees have previewed them (e.g., a teacher memorizes the correct answers to a certification exam).

These examples illustrate that access to psychological tests needs to be limited. But limited to whom? The answer, it turns out, depends on the complexity of the specific test under consideration. Guidelines proposed many years ago by the American Psychological Association (**APA, 1953** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib34)) are still relevant today, even though they are not enforced by all publishers. The APA proposed that tests fall into three levels of complexity (Levels A, B, and C) that require different degrees of expertise from the examiner. Level A comprised simple paper-and-pencil tests that require minimal training. These can be used by responsible nonpsychologists such as educational administrators. Examples include group educational tests and vocational proficiency scales. Level B tests require training in statistics and knowledge of test construction. Some graduate training is needed. This group includes aptitude tests and personality inventories relevant to normal populations. Level C includes the most complex instruments. Minimum training required is a master's degree in psychology or a related field. Instruments include projective personality tests, individual tests of intelligence, and neuropsychological test batteries.

In general, test publishers try to screen out inappropriate requests by requiring that purchasers have the necessary credentials. For example, the Psychological Corporation, one of the major suppliers of test materials in the United States, requires prospective customers to fill out a registration form detailing their training and experience with tests. Buyers who do not hold an advanced degree in psychology must list details of courses in the administration and interpretation of tests and in statistics. References are required, too.

Most test publishers also specify that individuals or groups who provide testing and counseling by mail are not allowed to purchase materials. On a related note, ethical standards now discourage practitioners from giving "take-home" tests to clients. Until recent years, this has been an occasional practice with lengthy personality tests such as the MMPI. The ethics committee endorsed the following point:

> Nonmonitored administration of the MMPI generally does not represent sound testing practice and may result in invalid assessment for a variety of reasons (e.g., influence from other people or completion of the test while intoxicated).

In general, users are advised to refrain from giving take-home tests and publishers are counseled to deny access to practitioners or groups who promote this practice.

Even though publishers attempt to filter out unqualified purchasers, there may still be instances in which sensitive tests are sold to unscrupulous individuals. Oles and Davis (**1977** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1252)) discovered that graduate students in psychology could purchase the WISC-R, MMPI, TAT, Stanford-Binet, and 16PF if they typed their orders on college stationery, placed the letters *Ph.D.* after their names, enclosed payment, and used a post office box return address. Although illicit test orders are few in number, they do occur.

1.13 RESPONSIBILITIES OF TEST USERS

The psychological assessment of personality, interests, brain functioning, aptitude, or intelligence is a sensitive professional action that should be completed with utmost concern for the well-being of the examinee, his or her family, employers, and the wider network of social institutions that might be affected by the results of that particular clinical assessment (**Matarazzo, 1990** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1059)). Over the years, the profession of psychology has proposed, clarified, and sharpened a series of thorough and thoughtful standards to provide guidance for the individual practitioner. Professional organizations publish formal ethical principles that bear upon test use, including the American Psychological Association (**APA, 2002** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib41)), the American Association for Counseling and Development (**AACD, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib28)), the American Speech-Language-Hearing Association (**ASHA, 1991** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib43)), and the National Association of School Psychologists (**NASP, 2010** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1215)).

In addition to ethical principles, several testing organizations have published practice guidelines to help define the scope of responsible test use. Sources of test use guidelines include teaching groups (**AFT, NCME, NEA, 1990** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib31)), the American Psychological Association (**APA, 1992b** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib38)), the Educational Testing Service (**ETS, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib449)), the Joint Committee on Testing Practices (**JCTP, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib846)), the Society for Industrial and Organizational Psychology (**SIOP, 1987** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1519)), and professional alliances (**AERA, APA, NCME, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30)). Finally, we should mention that the principles of responsible test use have been distilled in an illuminating casebook published jointly by several testing groups (**Eyde, Robertson, & Krug, 2009** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib473)).

The dozens of guidelines relevant to testing are quite specific, for example:

> Standard 5.9: When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Because of their specificity, a detailed analysis of relevant ethical and professional standards is beyond the scope of this text. What follows is a summary of the general provisions that pertain to the responsible practice of psychological testing and clinical psychological assessment.

These principles apply to psychologists, students of psychology, and others who work under the supervision of a psychologist. We restrict our discussion to those principles that are directly pertinent to the practice of psychological testing. Proper adherence to these principles would eliminate most—but not all—legal challenges to testing.

Best Interests of the Client

Several ethical principles recognize that all psychological services, including assessment, are provided within the context of a professional relationship. Psychologists are, therefore, enjoined to accept the responsibility implicit in this relationship. In general, the practitioner is guided by one overriding question: What is in the best interests of the client? The functional implication of this guideline is that assessment should serve a constructive purpose for the individual examinee. If it does not, the practitioner is probably violating one or more specific ethical principles. For example, Standard 11.15 in the *Standards* manual (**AERA, APA, NCME, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30)) warns testers to avoid actions that have unintended negative consequences. Allowing a client to attach unsupported surplus meanings to test results would not be in the best interests of the client and would, therefore, constitute an unethical testing practice. In fact, with certain worry-prone and self-doubting clients, a psychologist may choose not to use an appropriate test, since these clients are almost certain to engage in self-destructive misinterpretation of virtually any test findings.

Confidentiality and the Duty to Warn

Practitioners have a primary obligation to safeguard the confidentiality of information, including test results, that they obtain from clients in the course of consultations (Principle 5; **APA, 1992a** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib37)). Such information can be ethically released to others only after the client or a legal representative gives unambiguous consent, usually in written form. The only exceptions to confidentiality involve those unusual circumstances in which the withholding of information would present a clear danger to the client or other persons. For example, most states have passed laws that mandate that health care practitioners must report all cases of suspected abuse in children and vulnerable elderly persons. In most states, a psychologist who learns in the course of testing that the client has physically or sexually abused a child is obligated to report that information to law enforcement.

Psychologists also have a **duty to warn** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss96) that stems from the 1976 decision in the *Tarasoff* case (**Wrightsman, Nietzel, Fortune, & Greene, 2002** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1790)). Tanya Tarasoff was a young college student in California who was murdered by Prosenjit Poddar, a student from India. What makes the case relevant to the practice of psychology is that Poddar had made death threats regarding Tarasoff to his campus-based therapist. Although the therapist warned the police that Poddar had made death threats, he did not warn Tarasoff. Two months later, Poddar stabbed Tarasoff to death at her home. The parents of Tanya Tarasoff sued, and the California Supreme Court later agreed that therapists have a duty to use "reasonable care" to protect potential victims from their clients. Although the Tarasoff ruling has been modified by legislation in many states, the thrust of the case still stands: Clinicians must communicate any serious threat to the potential victim, law enforcement agencies, or both.

Finally, the clinician should consider the client's welfare in deciding whether to release information, especially when the client is a minor who is unable to give voluntary, informed consent. When appropriate, practitioners are advised to inform their clients of the legal limits of confidentiality.

Expertise of the Test User

A number of principles acknowledge that the test user must accept ultimate responsibility for the proper application of tests. From a practical standpoint, this means that the test user must be well trained in assessment and measurement theory. The user must possess the expertise needed to evaluate psychological tests for proper standardization, reliability, validity, interpretive accuracy, and other psychometric characteristics. This guideline has special significance in areas such as job screening, special education, testing of persons with disabilities, or other situations in which potential impact is strong.

Psychologists who are poorly trained in their chosen instruments can make serious errors of test interpretation that harm examinees. Furthermore, inept test usage may expose the examiner to professional sanctions and civil lawsuits. A common error observed among inexperienced test users is the overzealous, pathologized interpretation of personality test results (**Case Exhibit 1.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec13#ch01exh3)).

CASE EXHIBIT 1.3

Overzealous Interpretation of the MMPI

An inexperienced consulting psychologist routinely used the MMPI for preemployment screening of law enforcement candidates. One candidate subsequently filed a lawsuit, alleging that she had been harmed by the psychologist's report. The plaintiff, a young woman with extensive training and background in law enforcement, was denied a position as police officer because of a supposedly "defensive" MMPI profile. Her profile was entirely within normal limits, although she did obtain a *T* score of 72 on the K scale. The K scale is usually considered a good index of defensive test-taking attitudes, especially for mental health evaluations with clinic or hospital referrals. By way of quick review, MMPI *T* scores of approximately 50 are average, whereas elevations of 70 or higher are considered noteworthy. The consulting psychologist noticed the candidate's elevated score on the K scale, surmised hastily that the candidate was unduly defensive, and cautioned the police chief not to hire her.

What the psychologist did not know is that elevated K-scale scores are extremely common among law enforcement job applicants. For example, Hiatt and Hargrave (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib742)) found that about 25 percent of a sample of peace officers produced MMPI profiles with K scales at or above a *T* score of 70. In fact, successful police officers tend to have higher K-scale scores than "problem" peace officers! In this case the test user did not possess sufficient expertise to use the MMPI for job screening. His ignorance on this point constituted a breach of professional ethics. Incidentally, the case was settled out of court for a substantial sum of money, showing that trespasses of responsible test use can have serious legal consequences.

The expertise of the psychologist is particularly relevant when test scoring and interpretation services are used. The Ethical Principles of the American Psychological Association leave no room for doubt:

> Psychologists retain appropriate responsibility for the appropriate application, interpretation, and use of assessment instruments, whether they score and interpret such tests themselves or use automated or other services. (**APA, 1992a** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib37))

The reader is referred to **Topic 12B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12box3) , Computerized Assessment and the Future of Testing, for further discussion of this point.

Informed Consent

Before testing commences, the test user needs to obtain informed consent from test takers or their legal representatives. Exceptions to informed consent can be made in certain instances, for example, legally mandated statewide testing programs, school-based group testing, and when consent is clearly implied (e.g., college admissions testing). The principle of **informed consent** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss156) is so important that the *Standards* manual devotes a separate standard to it:

> Informed consent implies that the test takers or representatives are made aware, in language that they can understand, of the reasons for testing, the type of tests to be used, the intended use and the range of material consequences of the intended use. If written, video, or audio records are made of the testing session, or other records are kept, test takers are entitled to know what testing information will be released and to whom. (**AERA et al., 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib30))

Even young children or test takers with limited intelligence deserve an explanation of the reasons for assessment. For example, the examiner might explain, "I'm going to ask you some questions and have you work on some puzzles so I can see what you can do and find out what things you need more help with."

From a legal standpoint, the three elements of informed consent include disclosure, competency, and voluntariness (**Melton, Petrila, Poythress, & Slobogin, 1998** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1128)). The heart of disclosure is that the client receive sufficient information (e.g., about risks, benefits, release of reports) to make a thoughtful decision about continued participation in the testing. Competency refers to the mental capacity of the examinee to provide consent. In general, there is a presumption of competency unless the examinee is a child, very elderly, or has mental disabilities (e.g., has mental retardation). In these cases, a guardian will need to provide legal consent. Finally, the standard of voluntariness implies that the choice to undergo an assessment battery is given freely and not based on subtle coercion (e.g., inmates are promised release time if they participate in research testing). In most cases, the examiner uses a written informed consent form such as that found in **Figure 1.5** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec13#ch01fig5) .

INFORMED CONSENT FOR PSYCHOLOGICAL ASSESSMENT

This is an agreement between [Client's Name] and Dr. [Practitioner's Name], a licensed psychologist in the state of Illinois. You are encouraged to ask questions at any time about my training and background, and about the process of testing.

1. ***General Information:*** The purpose of this assessment is to provide you (and possibly others) with information about your psychological functioning that could prove helpful. The assessment will involve a brief interview and psychological testing. The entire process will take about three to four hours.
2. ***Specific Procedures:*** In addition to interview, the following tests will be administered: [List of tests and brief descriptions], e.g., MMPI-2, a 567-item true-false inventory of psychological functioning. WAIS-IV, a general test of adult intelligence in varied areas.
3. ***Test Report:*** The relevant information from the interview and the test results will be summarized in a written report. The results and the report will be reviewed with you in approximately one week. I will keep a copy of this report in a locked file for at least seven years.
4. ***Confidentiality:*** The report will not be released to any other source unless you sign a formal request. A few (remote) exceptions to the confidentiality guideline include situations of potential harm to self or others, abuse of children or elderly, or a court order to release the test results.
5. ***Cost:*** An hourly rate of $____ is used in determining the total fee. I will bill your insurance company, but you are responsible for the cost. The estimated total cost for your assessment is $____.
6. ***Side Effects:*** While most people find these tests and procedures to be interesting, some people experience anxiety when tested. Yet, it is unlikely that you will experience any long-term adverse effects from this assessment. You are encouraged to talk about the experience as we proceed.
7. ***Refusal of Assessment:*** Most people find the process of psychological assessment to be beneficial. However, you are not required to undergo this assessment. You can withdraw consent and discontinue at any time. On request, I will discuss referral options with you.

_____ _____

Client's Signature Date

FIGURE 1.5 Abbreviated Example of Informed Consent for Psychological Assessment

Note: This form is illustrative only. Practitioners should consult legal counsel in regard to the details of an informed consent form.

Obsolete Tests and the Standard of Care

Standard of care is a loose concept that often arises in the professional or legal review of specific health practices, including psychological testing. The prevailing **standard of care** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss307) is one that is "usual, customary or reasonable" (**Rinas & Clyne-Jackson, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1369)). To cite an extreme example, in medicine the standard of care for a fever might include the administration of aspirin—but would not include the antiquated practice of bleeding the patient.

Practitioners of psychological testing must be wary of obsolete tests, because their use might violate the prevailing standard of care. A case in point is the MMPI versus the MMPI-2. Even though the MMPI-2 is a relatively conservative revision of the highly esteemed MMPI, the improvements in norming and scale construction are substantial. The MMPI-2 is now the standard of care in MMPI-based assessment of psychopathology. Practitioners who continue to rely on the original MMPI could be liable for malpractice suits, especially if the test interpretation resulted in misleading interpretive statements or an incorrect diagnosis.

Another concern relevant to the standard of care is reliance on test results that are outdated for the current purpose. After all, individual characteristics and traits show valid change over time. A student who meets the criteria for learning disability (LD) in the fourth grade might show large gains in academic achievement, such that the LD diagnosis is no longer accurate in the fifth grade. Personality test results are especially prone to quixotic change. A short-term personal crisis might cause an MMPI-2 profile to look like a range of mountains. A week later, the test profile could be completely normal. It is difficult to provide comprehensive guidelines as to the "shelf life" of psychological test results. For example, GRE test scores that are years old still might be validly predictive of performance in graduate school, whereas Beck Depression Inventory test results from yesterday could mislead a therapist as to the current level of depression. Practitioners must evaluate the need for retesting on an individual basis.

Responsible Report Writing

Except for group testing, the practice of psychological testing invariably culminates in a written report that constitutes a semipermanent record of test findings and examiner recommendations. Effective report writing is an important skill because of the potential lasting impact of the written document. It is beyond the scope of this text to illuminate the qualities of effective report writing, although we can refer the reader to a few sources (**Gregory, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib650) ; **Tallent, 1993** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1608)).

Responsible reports typically use simple and direct writing that steers clear of jargon and technical terms. The proper goal of a report is to provide helpful perspectives on the client, not to impress the referral source that the examiner is a learned person! When Tallent (**1993** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1608)) surveyed more than one thousand health practitioners who made referrals for testing, one respondent declared his disdain toward psychologists who "reflect their needs to shine as a psychoanalytic beacon in revealing the dark, deep secrets they have observed." On a related note, effective reports stay within the bounds of expertise of the examiner. For example:

> It is never appropriate for a psychologist to recommend that a client undergo a specific medical procedure (such as a CT scan for an apparent brain tumor) or receive a particular drug (such as Prozac for depression). Even when the need for a special procedure seems obvious (e.g., the symptoms strongly attest to the rapid onset of a brain disease), the best way to meet the needs of the client is to recommend immediate consultation with the appropriate medical profession (e.g., neurology or psychiatry). (**Gregory, 1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib650))

Additional advice on effective report writing can be found in Ownby (**1991** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1263**)) and Sattler (**2001** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1438**)).

Communication of Test Results

Individuals who take psychological tests anticipate that the results will be shared with them. Yet practitioners often do not include one-to-one feedback as part of the assessment. A major reason for reluctance is a lack of training in how to provide feedback, especially when the test results appear to be negative. For example, how does a clinician tell a college student that her IQ is 93 when most students in that milieu score 115 or higher?

Providing effective and constructive feedback to clients about their test results is a challenging skill to learn. Pope (**1992** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1314**)) emphasizes the responsibility of the clinician to determine that the client has understood adequately and accurately the information that the clinician was attempting to convey. Furthermore, it is the responsibility of the clinician to check for adverse reactions:

> Is the client exceptionally depressed by the findings? Is the client inferring from findings suggesting a learning disorder that the client—as the client has always suspected—is "stupid"? Using scrupulous care to conduct this assessment of the client's understanding of and reactions to the feedback is no less important than using adequate care in administering standardized psychological tests; test administration and feedback are equally important, fundamental aspects of the assessment process. (p. 271)

Proper and effective feedback involves give-and-take dialogue in which the clinician ascertains how the client has perceived the information and seeks to correct potentially harmful interpretations.

Destructive feedback often arises when the clinician fails to challenge a client's incorrect perceptions about the meaning of test results. Consider IQ tests in particular—a case in which many persons deify test scores and consider them an index of personal worth. Prior to providing test results, a clinician is advised to investigate the client's understanding of what IQ scores mean. After all, IQ is a limited slice of intellectual functioning: It does not evaluate drive or character of any kind, it is accurate only to about ±5 points, it may change over time, and it does not assess many important attributes such as creativity, social intelligence, musical ability, or athletic skill. But a client may have an unrealistic perspective about IQ and, hence, might jump to erroneous conclusions when hearing that her score is "only" 93. The careful practitioner will elicit the client's views and challenge them when needed before proceeding. Further thoughts on feedback can be found in Pope (**1992** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1314**)).

Going beyond the general pronouncement to avoid harm when providing test feedback, Finn and Tonsager (**1997** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib502**)) present the intriguing view that information about test results should be directly and immediately therapeutic to individuals experiencing psychological problems. In other words, they propose that psychological assessment is a form of short-term intervention, not just a basis for gathering information that is *later* used for therapeutic purposes. In one study (**Finn & Tonsager, 1992** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib500**)), they examined the effects of a brief psychological assessment on clients at a university counseling center. Thirty-two students took part in an initial interview, completed the MMPI-2, and then received a one-hour feedback session conducted according to a method developed by Finn (**1996** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib499**)). A comparison group of 29 students was interviewed and received an equal amount of supportive, nondirective psychotherapy instead of the test feedback. The clients in the MMPI-2 assessment group showed a greater decline in symptomatic distress and a greater increase in self-esteem, immediately following their feedback session and also two weeks later, than the clients in the comparison group. The feedback group also felt more hopeful about their problems after the brief assessment. These findings illustrate the importance of providing thoughtful and constructive test feedback instead of rushing through a perfunctory review of the results.

Consideration of Individual Differences

Knowledge of and respect for individual differences is highlighted by all professional organizations that deal with psychological testing. The American Psychological Association lists this as one of six guiding principles:

> Principle D: Respect for People's Rights and Dignity . . . Psychologists are aware of cultural, individual, and role differences, including those due to age, gender, race, ethnicity, national origin, religion, sexual orientation, disability, language, and socio-economic status. Psychologists try to eliminate the effect on their work of biases based on those factors, and they do not knowingly participate in or condone unfair discriminatory practices. (**APA, 1992a** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib37**))

The relevance of this principle to psychological testing is that practitioners are expected to know when a test or interpretation may not be applicable because of factors such as age, gender, race, ethnicity, national origin, religion, sexual orientation, disability, language, and socioeconomic status. We can illustrate this point with a case study reported in Eyde et al. (**1993** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib474**)). A psychologist evaluated a 75-year-old man at the request of his wife, who had noticed memory problems. The psychologist administered a mental status examination and a prominent intelligence test. Performance on the mental status examination was normal, but standard scores on the intelligence test revealed a large discrepancy between verbal subtests and subtests measuring spatial ability and processing speed. The psychologist interpreted this pattern as indicating a deterioration of intellectual functioning in the husband. Unfortunately, this interpretation was based on faulty use of non-age-corrected standard scores. Also, the psychologist did not assess for depression, which is known to cause visuospatial performance to drop sharply (**Wolff & Gregory, 1992** (**http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1779**)). In fact, a series of further evaluations revealed that the husband was a perfectly healthy 75-year-old man. The psychologist failed to consider the relevance of the gentleman's age and emotional status when interpreting the intelligence test. This was a costly oversight that caused the client and his wife substantial unnecessary worry.

1.14 TESTING OF CULTURAL AND LINGUISTIC MINORITIES

Background and Historical Notes

Persons of ethnic minority descent (non-European origin) currently constitute about a third of the U.S. population, and it is estimated that they will comprise more than 50 percent within several decades. Yet the enterprise of testing is based almost entirely on the efforts of white psychologists who bring an Anglo-American viewpoint to their work. The suitability of existing tests for the evaluation of diverse populations cannot be taken for granted. The assessment of ethnic minority individuals raises important questions, especially when test results translate to placement decisions or other sensitive outcomes, as is commonly the case within educational institutions.

Unfortunately, the early pioneers in the testing movement largely ignored the impact of cultural background on test results. For example, in the 1920s Henry Goddard concluded that the intelligence of the average immigrant was alarmingly low, "perhaps of moron grade." Yet he downplayed the likelihood that language and cultural differences could explain the low test scores of immigrants. Goddard's role in the history of testing is discussed in the next chapter.

Perhaps as a rebound against these early methods, beginning in the 1930s psychologists displayed an increased sensitivity to cultural variables in the practice of testing. A shining example in this regard was Stanley Porteus, who undertook a wide-ranging investigation of the temperament and intelligence of Australian aboriginal peoples. Porteus (**1931** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1317)) used many traditional instruments (block designs, mazes, digit span), but to his credit he also devised an ecologically valid measure of intelligence for this group, namely, footprint recognition. Whereas the aboriginal examinees performed poorly on the Eurocentric tests, their ability to recognize photographed footprints was on a par with other racial groups studied. Even so, Porteus displayed an acute awareness that his procedures *still* might have handicapped the aboriginals:

> The photograph of a footprint is not the same as the footprint itself, and quite probably a number of cues that are made use of by the aboriginal tracker are absent from a photograph. The varying depths of parts of the foot impression are not visible in the photograph, and the individual peculiarities other than general shape and size of the footprint may not be brought out clearly. Hence we must expect that the aboriginal subjects would be under some disadvantage in matching these photographs of footprints, as against recognition of the footprints themselves. (pp. 399–400)

In a similar vein, DuBois (**1939** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib437)) found that Pueblo Indian children displayed superior ability on his specially devised horse drawing test of mental ability, whereas they performed less well on the mainstream Goodenough (**1926** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib616)) Draw-A-Man test. From these early studies onward, psychologists have maintained a keen interest in the impact of language and culture on the meaning of test results.

The Impact of Cultural Background on Test Results

Practitioners need to appreciate that the cultural background of examinees will impact the entire process of assessment. For this reason, Sattler (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1437)) advises assessment psychologists to approach their task from a pluralistic standpoint:

> Cultural groups may vary with respect to cultural values (stemming in part from cultural shock, discontinuity, or conflict); language and nuances in language style; views of life and death; roles of family members; problem-solving strategies; attitudes toward education, mental health, and mental illness; and stage of acculturation (the group may follow traditional values, accept the dominant group's values, or be at some point between the two). You should adopt a frame of reference that will enable you to understand how particular behaviors make sense within each culture. (p. 505)

For example, it is often noted that Native Americans display a distinctive conception of time, emphasizing *present-time* as opposed to the *future-time* orientation that is so powerfully formative in white, middle-class America (**Panigua, 1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1265)). A possible implication of this cultural difference is that time limits might not mean the same thing for a Native American child as for a child from the mainstream culture. Perhaps the minority child will disregard the subtest instructions and work at a careful, measured pace rather than seeking quick solutions. Of course, this child would then obtain a misleadingly low score on that measure.

While acknowledging the impact of cultural differences on testing, it is also important to avoid stereotypical overgeneralization. Culture is not monolithic. Every person is unique. Some Native Americans will exhibit a distinctive orientation to time but perhaps most will not. The challenge for the practitioner is to observe the clinical details of performance and to identify the culture-based nuances of behavior that help determine the test results.

An ingenious study by Moore (**1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1170)) powerfully illustrates the relevance of cultural background for understanding the test performance of ethnic minority examinees. She compared not only the intelligence test scores but also *the qualitative manner* of responding to test demands in two groups of adopted African American children. One group of 23 children had been transracially adopted into middle-class white families. The other group of 23 children had been intraracially adopted into middle-class African American families. All children were adopted prior to age 2 and the backgrounds of the adoptive families were similar in terms of education and social class. Thus, group difference in test scores and test behaviors could be attributed mainly to differences in cultural background arising from the fact that one group was adopted into African American families, the other adopted into white families. Testing and observations were completed by two female African American examiners who were "blind" to the purposes of the study. Tested at 7 to 10 years of age, the transracially adopted children scored an average IQ of 117 on the WISC compared to an average IQ of 104 for the traditionally adopted children. These IQ results were not remarkable, insofar as Scarr and Weinberg reported similar findings years before.

The surprising and informative outcome of the study was that the two groups of children showed very different *qualitative* behaviors during testing. As a group, the children with lower IQ scores (those adopted by African American families) were less likely to spontaneously elaborate on their work responses and more likely simply to refuse to respond when presented with a test demand. Moore (**1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1170)) offers the following interpretations:

> Children's tendency to spontaneously elaborate on their work responses may be a very important index of their level of involvement in task performance, strategies for problem solving, level of motivation to generate a correct response, and level of adjustment to the standardized test situation. . . . Although the terminal not-work response is treated as an incorrect response, it does not actually provide any empirical documentation of what the child does or does not know or of what the child can and cannot do. The only information available is that the child did not respond to the demand. (p. 322)

The essential lesson of this study is that culturally based differences in response style may function to conceal the underlying competence of some examinees. Cautious interpretation of test results is always advisable, but this is especially important for examinees from culturally or linguistically diverse backgrounds.

The influence of cultural factors is not limited to the test performance of children but extends to adults as well. Terrell, Terrell, and Taylor (**1981** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1631)) investigated the effects of racial trust/mistrust on the intelligence test scores of African American college students. They identified African American students with high and low levels of mistrust of whites. Using a 2 × 2 design, half of each group was then administered an individual intelligence test by a white examiner, the other half by an African American examiner. As predicted, the analysis of variance revealed no differences for the main effects of race of examiner (white versus African American) or level of mistrust (high versus low) (**Figure 1.6** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec14#ch01fig6)). But a substantial interaction was revealed; namely, the high-mistrust group with an African American examiner scored much better than the high-mistrust group with a white examiner (average IQs of 96 versus 86, respectively). Put simply, cultural mistrust among African Americans was associated with significantly lower IQ scores, but *only* when the examiner was white.

Further illustrating cultural influences, Steele (**1997** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1546)) has proposed a theory that societal stereotypes about groups influence the immediate intellectual performance and also the long-term identity development of individual group members. He has applied this theory both to women—when stereotypes affect their achievement in math and sciences—and to African Americans—when stereotypes apparently depress their performance on standardized tests. Here we discuss his research on stereotype threat with African American college students (**Steele & Aronson, 1995** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1547)).

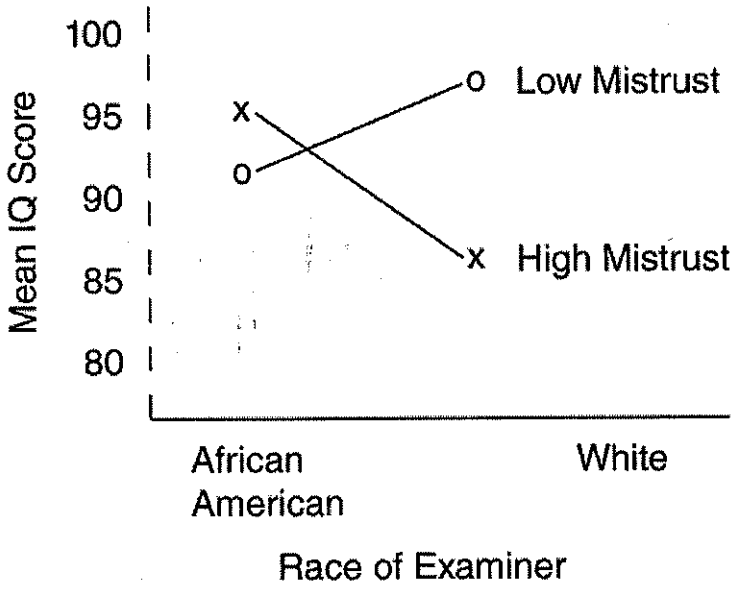


FIGURE 1.6 Mean IQ Scores of African American Students as a Function of Race of Examiner and Cultural Mistrust

Source: Based on data in Terrell, F., Terrell, S., & Taylor, J. (1981). Effects of race of examiner and cultural mistrust on the WAIS performance of Black students. *Journal of Consulting and Clinical Psychology, 49*, 750–751.

The idea of stereotype threat is essentially a sophisticated version of a self-fulfilling prophecy. The researchers define **stereotype threat** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss315) as the threat of confirming, as self-characteristic, a negative stereotype about one's group. For example, based on published data and media coverage about race and IQ scores, African Americans are stereotyped as possessing less intellectual ability than others. As a consequence, whenever they encounter tests of intelligence or academic achievement, individuals from this group may perceive a risk that they will confirm the stereotype. In the short run, stereotype threat is hypothesized to depress test performance through heightened anxiety and other mechanisms. In the long run, it may have the further impact of pressuring African American students to "protectively disidentify" with achievement in school and related intellectual domains.

Steele and Aronson (**1995** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1547)) conducted a series of four studies to evaluate the hypothesis of stereotype threat. All the investigations supported the hypothesis. We focus here on the first study, in which African American and white college students were given a 30-minute test composed of challenging items from the verbal Graduate Record Examination. Students from both racial groups were randomly assigned to one of three test conditions: stereotype-threat, in which the test was described as diagnostic of individual verbal ability; control, in which the test was described as a research tool only; and control-challenge, in which the test was described as a research tool only but participants were exhorted to "take this challenge seriously." Scores on the verbal test were adjusted (covariate analysis) on the basis of prior achievement scores so as to eliminate the effects of preexisting differences between groups.

Race differences were small and nonsignificant in the control and control-challenge conditions, whereas African Americans scored much lower than whites in the stereotype-threat condition (**Figure 1.7** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch01lev1sec14#ch01fig7)). In other studies, Steele and Aronson (**1995** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1547)) investigated the mechanism of mediation by which stereotype threat caused African Americans to score lower on standardized tests. The details are beyond the scope of this text, but the overall conclusion is not:

> Our best assessment is that stereotype threat caused an inefficiency of processing much like that caused by other evaluative pressures. Stereotype-threatened participants spent more time doing fewer items more inaccurately—probably as a result of alternating their attention between trying to answer the items and trying to assess the self-significance of their frustration. (**Steele & Aronson, 1995** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1547), p. 809)

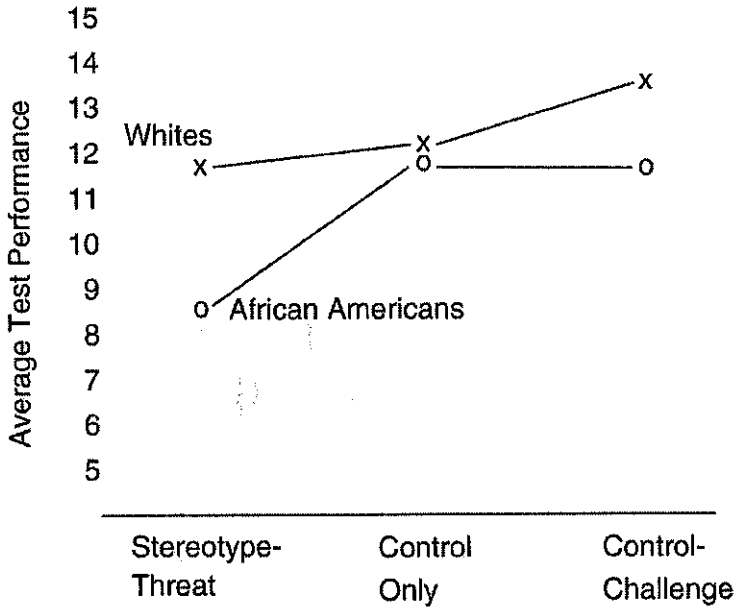


FIGURE 1.7 Average Verbal Items Correct for Whites and African Americans under Three Conditions

Source: Based on data in Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

In sum, the authors propose a social-psychological perspective on the meaning of lower test scores in African Americans and perhaps other stereotype-threatened groups as well. Their viewpoint emphasizes that test results do not reside within individuals. Test scores occur within a complex social-psychological field that is potentially influenced by national history, predicaments of race, and many other subtle factors.

1.15 UNINTENDED EFFECTS OF HIGH-STAKES TESTING

The prevailing view in the general public is that cheating rarely or never occurs in nationally administered testing programs. We tend to think that the risks are too high and the opportunities too limited for cheaters to prevail. Therefore, we rest assured that test fraud must be a rare event. Unfortunately, this view is probably naive. After all, a growing number of people must pass a test to gain college entry, get a job, or obtain a promotion. Furthermore, school officials increasingly are evaluated on the basis of average test scores in their district. Precisely because the stakes are so high, unscrupulous individuals will try to beat the system.

Widespread cheating in public school systems is sporadically reported in many large cities across the United States. In most cases, the cheating is motivated by the desire of teachers and principals to further their own careers by creating the illusion of educational excellence. For example, in 1999, dozens of teachers and two principals in the New York City public school system were charged with helping students cheat on the standardized reading and math tests used to rank schools and determine whether students move on to the next grade (*New York Times*, December 12, 1999). The cheating scheme was described as "one of the largest in the recent history of American public schools." In 2000, an entire eighth-grade class in a Chicago elementary school was required to retake the Iowa Tests of Basic Skills (ITBS) because a school administrator allegedly filled in incomplete tests and changed incorrect answers to correct ones (*Chicago Tribune*, June 2, 2000). Officials were tipped off to the fraud because the test scores were simply too good to be true—the *average* score for the class was two years above their standing. In 2005, the *Dallas Morning News* reported strong evidence of "organized, educator-led cheating" in dozens of schools on the statewide achievement test and found suspicious scores in *hundreds* more (**www.dallasnews.com** (http://www.dallasnews.com), March 21, 2005). Disturbingly, one assessment expert noted, "You're catching the dumb cheaters. The smart cheaters you're not going to be able to detect." We only read about the cases of cheating that are detected. The number of undetected cases is simply unknown, although probably larger than the public would like to believe.

Cheating in public school systems is not a thing of the past. It continues unabated, year after year. In 2011, a decade long cheating scandal was revealed in the Atlanta, Georgia, public school system (*Atlanta Journal-Constitution*, July 6, 2011). Teachers and principals routinely changed students' answer sheets to produce higher scores. The school system scores soared dramatically, bringing national acclaim to the district and the superintendent. But it was all based on fraud perpetrated by 178 educators, including 38 principals. Cheating was confirmed in 44 of 56 schools examined. In 2011, six charter schools in Los Angeles were threatened with closure when it was discovered that the founding director had ordered principals to open the state standardized tests and train students on actual test questions (*Los Angeles Times*, June 22, 2011). Suspiciously, scores for the schools had vaulted upward in recent years. The director and the six principals were terminated.

An especially flagrant instance of cheating on national tests was uncovered in Louisiana in 1997. This case involved wholesale circulation of the Educational Testing Service (ETS) exam administered to teachers who want to be school principals. As reported in the *New York Times* (September 28, 1997), copies of the 145-item test, along with correct answers, had circulated among teachers throughout southern Louisiana, most likely for several years. In a state ranked at or near the bottom on nearly every educational index, it appears that many potentially unqualified persons cheated their way into running the schools. ETS handled this case quietly by asking more than 200 teachers to retake the test so as to "confirm" their initial scores. Unfortunately, the Louisiana case was not an isolated instance. In another case, ETS allegedly failed to monitor its handling of the federal government's test for immigrants who want to become citizens, with the likely result that test supervisors accepted bribes. English-proficiency tests for foreign students also were vulnerable to cheating. In 1994, ETS canceled the scores of 30,000 students from China after discovering a ring that was selling the examinations abroad. Cizek (**1999** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib307)) catalogues literally dozens of ingenious ways that students have developed for cheating on tests: writing information on the floor, in tissues, on the back of a bottled water label; using an ultraviolet pen to write information on "blank" paper; and using a video transmitter (e.g., hidden in an eyeglass case) to send pictures of the test to an outside accomplice who then coaches the student by means of an audio receiver (e.g., hidden in the ear).

Stories about miniature transmitters are not fanciful. Consider the following story reported from a monolithic culture where test results literally make or break a child's future. In China, 10 million 18-year-olds take a two day exam each year that determines whether they will be allowed to attend public universities. Success or failure drastically impacts their lives and those of their families who might depend on their future income. In 2009, eight parents were jailed for up to three years after it was determined that they were transmitting stolen test answers to their children through miniature earpieces. The subterfuge was discovered when police detected unusual radio signals near the school (**www.guardian.co.uk** (http://www.guardian.co.uk), April 3, 2009).

In 2012, cheating was brought to light on the board certification test for radiology (CNN, *Prescription for Cheating*, January 13, 2012). For years, doctors around the country have helped one another cheat by each memorizing one or two test questions verbatim, writing down the questions after taking the test, and circulating the ever-expanding list of questions (dubbed "recalls") to cooperating programs. The practice is so widespread and considered so egregious that the American Board of Radiology released a sternly worded video condemning the use of recalls as unethical. CNN found at least 15 years' worth of test questions (with answers) on a website for residents in radiology.

Recently, efforts to circumvent exam security have become even more brazen, with some test preparation companies encouraging students to *steal* copies of college entrance exams such as the Scholastic Assessment Tests (SAT) (*Los Angeles Times*, October 12, 2005). Fortunately, the publisher of the SAT was granted a restraining order in federal court, prohibiting individuals or companies from soliciting stolen copies of the test. Even so, this episode illustrates once again that high-stakes testing has had a corrupting influence on the testing process.

Dishonest and inappropriate practices by school officials are implicated in the recent inflation of scores on nationally normed group tests of achievement. By definition, for a norm-referenced test, 50 percent of the examinees should score above the 50th percentile, 50 percent below. If the same test is used in a large sample of typical and representative school systems, average scores for the school systems should be split evenly—about half above the nationally normed 50th percentile, half below.

According to a survey reported in the news media (Foster, 1990), virtually all states of the union claim that average achievement scores for their school systems exceed the 50th percentile. The resulting overly optimistic picture of student achievement is labeled the **Lake Wobegon Effect** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss181), in reference to humorist Garrison Keillor's mythical Minnesota town where "all the children are above average."

How does inflation of achievement test scores arise? According to Cannell (**1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib267)), the major cause is educational administrators who are desperate to demonstrate the excellence of their school systems. Precisely because our society attaches so much importance to achievement test results, some educators apparently help students cheat on standardized tests. The alleged cheating includes the following:

- Teachers and principals coach students on test answers.

- Examiners give more than the allotted time to take tests.
- Administrators alter answer sheets.
- Teachers teach directly to the specific test items.
- Teachers make copies of the tests to give to their students.

In sum, the importance that our society attaches to achievement test scores has caused a number of unappealing side effects that undermine the very foundations of nationally normed group-testing programs.

Moore (**1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1172)**) reports on a special case in educational testing, namely, the districtwide consequences of court-ordered achievement testing. He surveyed 79 teachers from third- through fifth-grade level in a midwestern town in which the court required the use of a standardized test to determine the effectiveness of a desegregation effort. The test in question, the Iowa Tests of Basic Skills (ITBS), is a well-respected group achievement test that requires strict adherence to instructions and time limits for obtaining valid results. Yet the teachers found little value in the testing program, complaining that its benefits did not offset the time and costs involved. As a consequence of their devaluing the effort, nonstandard testing was practically the rule rather than the exception. The teachers engaged in several nonstandard practices, most of which tended to inflate the test scores. Inappropriate testing practices included praising students who answered a question correctly during the test (67 percent), using last year's test questions for practice (44 percent), recoding a student's answer sheet because he or she just "miscoded" the answer (26 percent), giving students as much time as they needed (24 percent), giving students items that were directly off the test (24 percent), and giving hints or clues during the test (23 percent). In general, Moore (**1994 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1172)**) notes that teachers modified their instructional efforts and curriculum in anticipation of having their students take the test. More than 90 percent of the teachers added test-related lessons to the curriculum, and more than 70 percent eliminated topics so that they could spend more time on test-related skills.

What this study demonstrates is that mandated educational testing can have the unanticipated consequence of polluting the validity of a worthy test—especially when crucial stakeholders have no voice in the process.

Further, in teaching to the tests, educators may emphasize bits and pieces of factual knowledge rather than imparting a general ability to think clearly and solve problems. In conclusion, it appears that an excessive emphasis on nationally normed achievement tests for selection and evaluation promotes inappropriate behavior, including outright fraud and cheating on the part of students and school officials. Just how widespread is the problem? Although we live with the optimistic assumption that fraud in nationally normed testing programs is rare, the disturbing truth is that we really don't know how often this occurs.

1.16 REPRISE: RESPONSIBLE TEST USE

We return now to the real-life quandaries of testing mentioned at the beginning of the topic. The reader will recall that the first quandary had to do with whether a consulting psychologist responsibly could refuse to provide feedback to police officer candidates referred for preemployment screening. Surprisingly, the answer to this query is "Yes." Under normal circumstances, a practitioner must explain assessment results to the client. But there are exceptions, as explained by Principle 9.10 of the APA Ethical Code:

> Psychologists take reasonable steps to ensure that explanations of results are given to the individual or designated representative unless the nature of the relationship precludes provision of an explanation of results (such as in some organizational consulting, preemployment or security screenings, and forensic evaluations), and this fact has been clearly explained to the person being assessed in advance.

The second quandary concerned a counselor who continued to use the MMPI even though the MMPI-2 has been available for several years. Is the counselor's refusal to use the MMPI-2 a breach of professional standards? The answer to this query is probably "Yes." The MMPI-2 is well validated and constitutes a significant improvement upon the MMPI. As mentioned previously, the MMPI-2 is now the standard of care in MMPI-based assessment of psychopathology. The counselor who continued to rely on the original MMPI could be liable for malpractice suits, especially if his test interpretations resulted in misleading interpretive statements or a false diagnosis.

The third predicament involved the use of a neighborhood friend as translator in the administration of the WISC-IV to a 9-year-old boy whose first language was Spanish. This is usually a mistake as it sacrifices strict control of the testing material. The examiner was not bilingual and, therefore, he would have no way of knowing whether the translator was remaining faithful to the original text or was possibly supplying additional cues. In an ideal world, the proper procedure would be to enlist a Spanish-speaking examiner who would use a test formally translated and also standardized with Hispanic examinees. For example, the Escala de Inteligencia Wechsler Para Ninos-Revisada de Puerto Rico (EIWN-R PR) would be a good choice.

The final quandary concerned the client who informed a psychologist that her recently deceased brother was most likely a pedophile. Is the psychologist obligated to report this case to law enforcement? The answer to this query is probably "Yes," but it may depend on the jurisdiction of the psychologist and the wording of the relevant statutes. In fact, the psychologist did report the case to authorities with unexpected consequences. Police obtained a search warrant, went to the home of the client's mother (where the brother had lived), and ransacked the brother's bedroom. The mother was traumatized by the unexpected visit from the police and blamed the fiasco on her daughter. A bitter estrangement followed, and the client then sued the psychologist for violation of confidentiality!

CHAPTER 3
Norms and Reliability

TOPIC 3A Norms and Test Standardization

3.1 Raw Scores (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec1#ch03lev1sec1)

3.2 Essential Statistical Concepts (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03lev1sec2)

3.3 Raw Score Transformations (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03lev1sec3)

3.4 Selecting a Norm Group (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec4#ch03lev1sec4)

3.5 Criterion-Referenced Tests (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec5#ch03lev1sec5)

This chapter concerns two basic concepts needed to facilitate the examiner's interpretation of test scores: norms and reliability. In most cases, scores on psychological tests are interpreted by reference to norms that are based on the distribution of scores obtained by a representative sample of examinees. In **Topic 3A** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03#ch03box1), Norms and Test Standardization, we review the process of standardizing a test against an appropriate norm group so that test users can make sense out of individual test scores. Since the utility of a test score is also determined by the consistency or repeatability of test results, we introduce the essentials of reliability theory and measurement in **Topic 3B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec5#ch03box2), Concepts of Reliability. The next chapter flows logically from the material presented here and investigates the complex issues of validity—does a test measure what it is supposed to measure? First, we begin with the more straightforward issues of establishing a comparative frame of reference (norms) and determining the consistency or repeatability of test results (reliability).

The initial outcome of testing is typically a raw score such as the total number of personality statements endorsed in a particular direction or the total number of problems solved correctly, perhaps with bonus points added in for quick solutions. In most cases, the initial score is useless by itself. For test results to be meaningful, examiners must be able to convert the initial score to some form of derived score based on comparison to a standardization or norm group. The vast majority of tests are interpreted by comparing individual results to a norm group performance; criterion-referenced tests are an exception, discussed subsequently.

A **norm group** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss218) consists of a sample of examinees who are representative of the population for whom the test is intended. Consider a word knowledge test designed for use with prospective first-year college students. In this case, the performance of a large, heterogeneous, nationwide sampling of such persons might be collected for purposes of standardization. The essential objective of test standardization is to determine the distribution of raw scores in the norm group so that the test developer can publish derived scores known as norms. Norms come in many varieties, for example, percentile ranks, age equivalents, grade equivalents, or standard scores, as discussed in the following. In general, norms indicate an examinee's standing on the test relative to the performance of other persons of the same age, grade, sex, and so on.

To be effective, norms must be obtained with great care and constructed according to well-known precepts discussed in the following. Furthermore, norms may become outmoded in just a few years, so periodic renorming of tests should be the rule, not the exception. We approach the topic of norms indirectly, first providing the reader with a discussion of raw scores and then reviewing statistical concepts essential to an understanding of norms.

3.1 RAW SCORES

The most basic level of information provided by a psychological test is the **raw score** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss270). For example, in personality testing, the raw score is often the number of questions answered in the keyed direction for a specific scale. In ability testing, the raw score commonly consists of the number of problems answered correctly, often with bonus points added for quick performance. Thus, the initial outcome of testing is almost always a numerical tally such as 17 out of 44 items answered in the keyed direction on a depression scale, or 29 of 55 raw score points earned on the block design subscale of an intelligence test.

However, it should be obvious to the reader that raw scores, in isolation, are absolutely meaningless. For example, what use is it to know that a subject correctly solved 12 of 20 abstract reasoning questions? What does it mean that an examinee responded in the keyed direction to 19 out of 33 true-false questions from a psychological-mindedness scale?

It is difficult to even think about such questions without resorting to comparisons of one variety or another. We want to know how others have done on these tests, whether the observed scores are high or low in comparison to a representative group of subjects. In the case of ability tests, we are curious whether the questions were easy or hard, especially in relation to the age of the subject.

In fact, it seems almost a truism that a raw score becomes meaningful mainly in relation to norms, an independently established frame of reference derived from a standardization sample. We have much to say about the derivation and use of norms later in this unit. For now it will suffice to know that norms are empirically established by administering a test to a large and representative sample of persons. An examinee's score is then compared to the distribution of scores obtained by the standardization sample. In this manner, we determine from the norms whether an obtained score is low, average, or high.

The vast majority of psychological tests are interpreted by consulting norms; as noted, these instruments are called *norm-referenced tests.* However, the reader is reminded that other kinds of instruments do exist. In particular, criterion-referenced tests help determine whether a person can accomplish an objectively defined criterion such as adding pairs of two-digit numbers with 97 percent accuracy. In the case of criterion-referenced tests, norms are not essential. We elaborate on criterion-referenced tests at the end of this topic.

There are many different kinds of norms, but they share one characteristic: Each incorporates a statistical summary of a large body of scores. Thus, in order to understand norms, the reader needs to master elementary descriptive statistics. We take a modest digression here to review essential statistical concepts.

3.2 ESSENTIAL STATISTICAL CONCEPTS

Suppose for the moment that we have access to a high-level vocabulary test appropriate for testing the verbal skills of college professors and other professional persons (**Gregory & Gernert, 1990** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib652)). The test is a multiple-choice quiz of 30 difficult words such as *welkin, halcyon,* and *mellifluous.* A curious professor takes the test and chooses the correct alternative for 17 of the 30 words. She asks how her score compares to others of similar academic standing. How might we respond to her question?

One manner of answering the query would be to give her a list of the raw scores from the preliminary standardization sample of 100 representative professors at her university (**Table 3.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab1)). However, even with this relatively small norm sample (thousands of subjects is more typical), the list of test scores is an overpowering display.

TABLE 3.1 Raw Scores of 100 Professors on a 30-Item Vocabulary Test

6,	10,	16,	16,	17,	14,	19,	14,	16,	15
17,	17,	19,	20,	20,	22,	17,	24,	14,	25
13,	20,	11,	20,	21,	11,	20,	16,	18,	12
13,	7,	20,	27,	21,	7,	15,	18,	18,	25
20,	27,	28,	13,	21,	17,	12,	18,	12,	15
9,	24,	25,	9,	17,	17,	9,	19,	24,	15
20,	21,	22,	12,	21,	12,	19,	19,	23,	16
8,	12,	12,	17,	13,	19,	13,	11,	16,	16
7,	19,	14,	17,	19,	14,	18,	15,	15,	15
14,	14,	17,	18,	18,	22,	11,	15,	13,	9

Source: Based on data from Gregory, R. J., & Gernert, C. H. (1990). *Age trends for fluid and crystallized intelligence in an able subpopulation.* Unpublished manuscript.

When confronted with a collection of quantitative data, the natural human tendency is to summarize, condense, and organize it into meaningful patterns. For example, in assessing the meaning of the curious professor's vocabulary score, the reader might calculate the average score for the entire sample, or tally the relative position of the professor's score (17 correct) among the 100 data points found in **Table 3.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab1) . We review these and other approaches to organizing and summarizing quantitative data in the following sections.

Frequency Distributions

A very simple and useful way of summarizing data is to tabulate a frequency distribution (**Table 3.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab2)). A **frequency distribution** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss126) is prepared by specifying a small number of usually equal-sized class intervals and then tallying how many scores fall within each interval. The sums of the frequencies for all intervals will equal *N*, the total number of scores in the sample. There is no hard and fast rule for determining the size of the intervals. Obviously, the size of the intervals depends on the number of intervals desired. It is common for frequency distributions to include between 5 and 15 class intervals. In the case of **Table 3.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab2) , there are 9 class intervals of 3 scores each. The table indicates that one professor scored 4, 5, or 6, eight professors scored 7, 8, or 9, and so on.

TABLE 3.2 Frequency Distribution of Scores of 100 Professors on a Vocabulary Test

Class Interval	*Frequency*
4–6	1
7–9	8
10–12	12
13–15	21
16–18	24
19–21	21
22–24	7
25–27	5
28–30	1
	N = 100

A **histogram** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss146) provides a graphic representation of the same information contained in the frequency distribution (**Figure 3.1a** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig1)). The horizontal axis portrays the scores grouped into class intervals, whereas the vertical axis depicts the number of scores falling within each class interval. In a histogram, the height of a column indicates the number of scores occurring within that interval. A **frequency polygon** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss127) is similar to a histogram, except that the frequency of the class intervals is represented by single points rather than columns. The single points are then joined by straight lines (**Figure 3.1b** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig1)).

The graphs shown in **Figure 3.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig1) constitute visual summaries of the 100 raw score data points from the sample of professors. In addition to visual summaries of data, it is also possible to produce numerical summaries by computing statistical indices of central tendency and dispersion.

Measures of Central Tendency

Can we designate a single, representative score for the 100 vocabulary scores in our sample? The mean (M), or arithmetic average, is one such measure of central tendency. We compute the **mean** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss192) by adding all the scores up and dividing by N, the number of scores. Another useful index of central tendency is the **median** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss194), the middlemost score when all the scores have been ranked. If the number of scores is even, the median is the average of the middlemost two scores. In either case, the median is the point that bisects the distribution so that half of the cases fall above it, half below. Finally, the **mode** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss207) is simply the most frequently occurring score. If two scores tie for highest frequency of occurrence, the distribution is said to be bimodal.

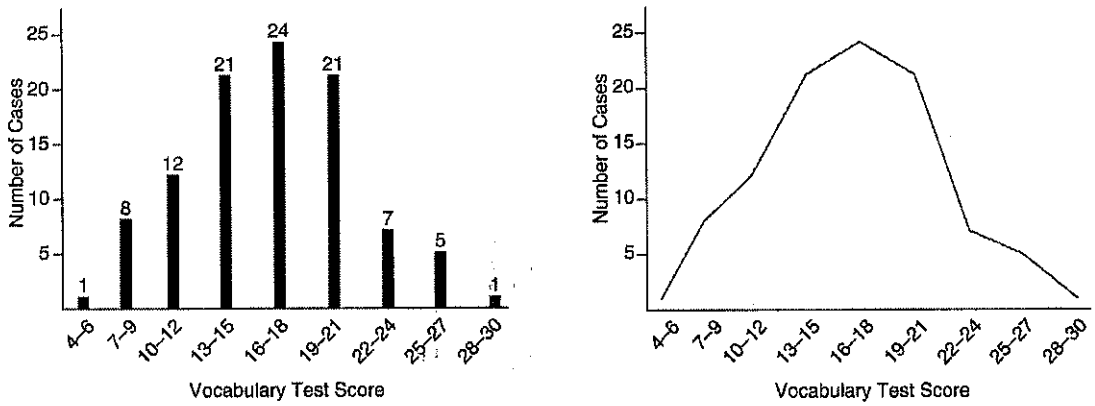


FIGURE 3.1 (a) A Histogram Representing Vocabulary Test Scores for 100 Professors. (b) A Frequency Polygon of Vocabulary Test Scores for 100 Professors

The mean of the scores listed in **Table 3.1** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab1) is 16.8; the median and mode are both 17. In this instance, the three measures of central tendency are in very good agreement. However, this is not always so. The mean is sensitive to extreme values and can be misleading if a distribution has a few scores that are unusually high or low. Consider an extreme case in which nine persons earn $10,000 and a tenth person earns $910,000. The mean income for this group is $100,000, yet this income level is not typical of anyone in the group. The median income of $10,000 is much more representative. Of course, this is an extreme example, but it illustrates a general point: If a distribution of scores is skewed (that is, asymmetrical), the median is a better index of central tendency than the mean.

Measures of Variability

Two or more distributions of test scores may have the same mean, yet differ greatly in the extent of dispersion of the scores about the mean (**Figure 3.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig2)). To describe the degree of dispersion, we need a statistical index that expresses the variability of scores in the distribution.

The most commonly used statistical index of variability in a group of scores is the **standard deviation** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss303), designated as s or abbreviated as SD. From a conceptual standpoint, the reader needs to know that the standard deviation reflects the degree of dispersion in a group of scores. If the scores are tightly packed around a central value, the standard deviation is small. In fact, in the extreme case in which all the scores are identical, the standard deviation is exactly zero. As a group of scores becomes more spread out, the standard deviation becomes larger. For example, in **Figure 3.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig2), distribution a would have the largest standard deviation, distribution c the smallest.

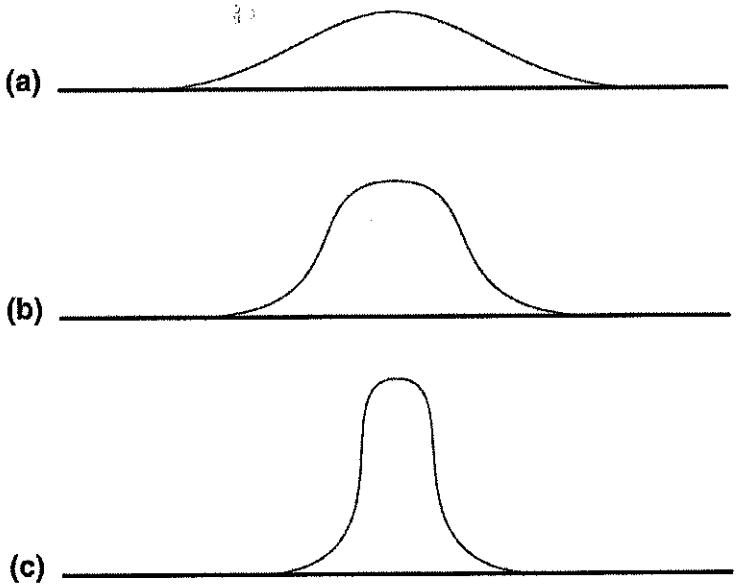


FIGURE 3.2 Three Distributions with Identical Means but Different Variability

The standard deviation, or *s*, is simply the square root of the variance, designated as s^2. The formula for the **variance** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss345) is

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{(N - 1)}$$

where Σ designates "the sum of," X stands for each individual score, $\overline{X}$ is the mean of the scores, and N is the total number of scores. As the name suggests, the variance is a measure of variability. However, psychologists usually prefer to report the standard deviation, which is computed by taking the square root of the variance. Of course, the variance and the standard deviation convey interchangeable information—one can be computed from the other by squaring (the standard deviation to obtain the variance) or taking the square root (of the variance to obtain the standard deviation). The standard deviation is nonetheless the preferred measure of variance in psychological testing because of its direct relevance to the normal distribution, as discussed in the next section.

The Normal Distribution

The frequency polygon depicted in **Figure 3.1b** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig1) is highly irregular in shape, a typical finding with real-world data based on small sample sizes. What would happen to the shape of the frequency polygon if we increased the size of the normative sample and also increased the number of class intervals by reducing their size? Possibly, as we added new subjects to our sample, the distribution of scores would more and more closely resemble a symmetrical, mathematically defined, bell-shaped curve called the **normal distribution** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss220) (**Figure 3.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig3)).

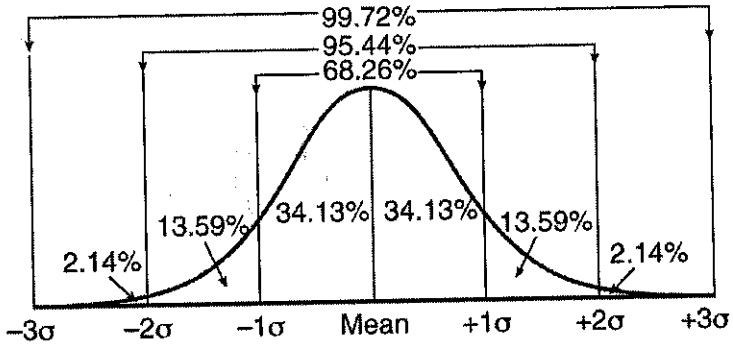


FIGURE 3.3 The Normal Curve and the Percentage of Cases within Certain Intervals

Psychologists prefer a normal distribution of test scores, even though many other distributions are theoretically possible. For example, a rectangular distribution of test scores—an equal number of outcomes in each class interval—is within the realm of possibility. Indeed, many laypersons might even prefer a rectangular distribution of test scores on the egalitarian premise that individual differences are thereby less pronounced. For example, a higher proportion of persons would score in the superior range if psychological tests conformed to a rectangular rather than normal distribution of scores.

Why, then, do psychologists prefer a normal distribution of test scores, even to the point of selecting test items that help produce this kind of distribution in the standardization sample? There are several reasons, including statistical considerations and empirical findings. We digress briefly here to explain the psychometric fascination with normal distributions.

One reason that psychologists prefer normal distributions is that the normal curve has useful mathematical features that form the basis for several kinds of statistical investigation. For example, suppose we wished to determine whether the average IQs for two groups of subjects were significantly different. An inferential statistic such as the *t*-test for a difference between means would be appropriate. However, many inferential statistics are based on the assumption that the underlying population of scores is normally distributed, or nearly so. Thus, in order to facilitate the use of inferential statistics, psychologists prefer that test scores in the general population follow a normal or near-normal distribution.

Another basis for preferring the normal distribution is its mathematical precision. Since the normal distribution is precisely defined in mathematical terms, it is possible to compute the area under different regions of the curve with great accuracy. Thus, a useful property of normal distributions is that the percentage of cases falling within a certain range or beyond a certain value is precisely known. For example, in a normal distribution, a mere 2.14 percent of the scores will exceed the mean by two standard deviations or more (**Figure 3.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig3)). In like manner, we can determine that the vast bulk of scores—more than 68 percent—fall within one standard deviation of the mean in either direction.

A third basis for preferring a normal distribution of test scores is that the normal curve often arises spontaneously in nature. In fact, early investigators were so impressed with the ubiquity of the normal distribution that they virtually deified the normal curve as a law of nature. For example, Galton (**1888** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib560)) wrote:

> It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

Certainly there is no "law of nature" regarding the form that frequency distributions must take. Nonetheless, it is true that many important human characteristics—both physical and mental—produce a close approximation to the normal curve when measurements for large and heterogeneous samples are graphed. For example, a near-normal distribution curve is a well-known finding for physical characteristics such as birthweight, height, and brain weight (**Jensen, 1980** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib832)). An approximately normal distribution is also found with numerous mental tests, even for tests constructed entirely without reference to the normal curve.

Skewness

Skewness (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss294) refers to the symmetry or asymmetry of a frequency distribution. If test scores are piled up at the low end of the scale, the distribution is said to be positively skewed. In the opposite case, when test scores are piled up at the high end of the scale, the distribution is said to be negatively skewed (**Figure 3.4 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig4)**).

In psychological testing, skewed distributions usually signify that the test developer has included too few easy items or too few hard items. For example, when scores in the standardization sample are massed at the low end (positive skew), the test probably contains too few easy items to make effective discriminations at this end of the scale. In this case, examinees who obtain zero or near-zero scores might actually differ with respect to the dimension measured. However, the test is unable to elicit these differences, since most of the items are too hard for these examinees. Of course, the opposite pattern holds as well. If scores are massed at the high end (negative skew), the test probably contains too few hard items to make effective discriminations at this end of the scale.

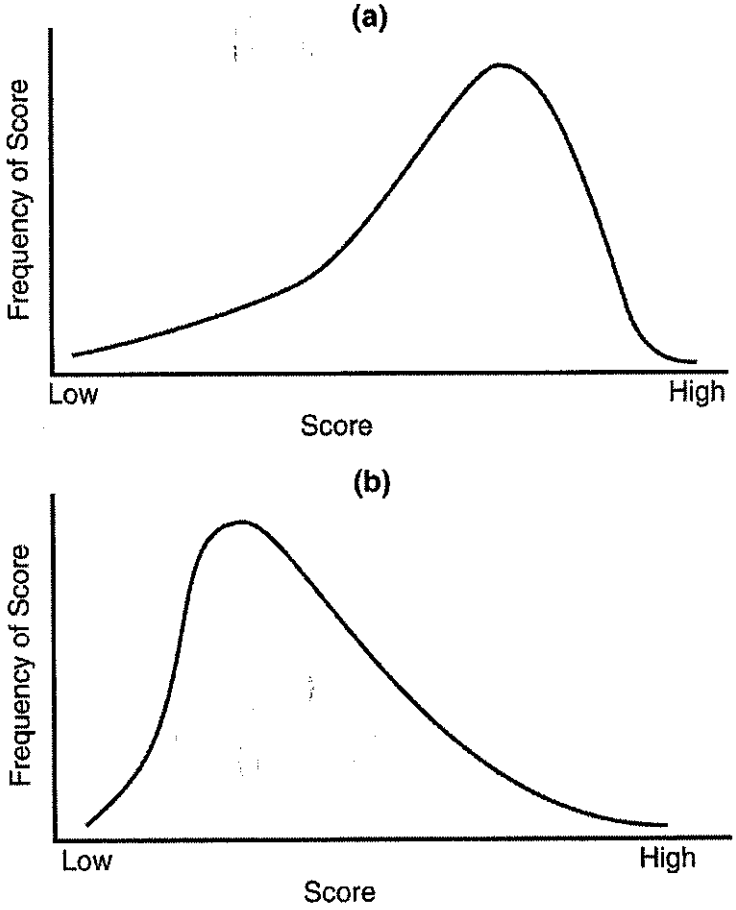


FIGURE 3.4 Skewed Distribution Curves: (a) Negative Skew; (b) Positive Skew

When initial research indicates that an instrument produces skewed results in the standardization sample, test developers typically revamp the test at the item level. The most straightforward solution is to add items or modify existing items so that the test has more easy items (to reduce positive skew) or more hard items (to reduce negative skew). If it is too late to revise the instrument, the test developer can use a statistical transformation to help produce a more normal distribution of scores (see the following). However, the preferred strategy is to revise the test so that skewness is minimal or nonexistent.

3.3 RAW SCORE TRANSFORMATIONS

Making sense out of test results is largely a matter of transforming the raw scores into more interpretable and useful forms of information. In the preceding discussion of normal distributions, we hinted at transformations by showing how knowledge of the mean and standard deviation of such distributions can help us determine the relative standing of an individual score. In this section we continue this theme in a more direct manner by introducing the formal requirements for several kinds of raw score transformations.

Percentiles and Percentile Ranks

A **percentile** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss236) expresses the percentage of persons in the standardization sample who scored below a specific raw score. For example, on the vocabulary test depicted in **Table 3.2** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03tab2) , 94 percent of the sample fell below a raw score of 25. Thus, a raw score of 25 would correspond to a percentile of 94, denoted as P_{94}. Note that higher percentiles indicate higher scores. In the extreme case, an examinee who obtained a raw score that exceeded every score in the standardization sample would receive a percentile of 100, or P_{100}.

The reader is warned not to confuse percentiles with percent correct. Remember that a percentile indicates only how an examinee compares to the standardization sample and does not convey the percentage of questions answered correctly. Conceivably, on a difficult test, a raw score of 50 percent correct might translate to a percentile of 90, 95, or even 100. Conversely, on an easy test, a raw score of 95 percent correct might translate to a percentile of 5, 10, or 20.

Percentiles can also be viewed as ranks in a group of 100 representative subjects, with 1 being the lowest rank and 100 the highest. Note that percentile ranks are the complete reverse of usual ranking procedures. A percentile rank (PR) of 1 is at the bottom of the sample, while a PR of 99 is near the top.

A percentile of 50 (P_{50}) corresponds to the median or middlemost raw score. A percentile of 25 (P_{25}) is often denoted as $Q1$ or the first quartile because one-quarter of the scores fall below this point. In like manner, a percentile of 75 (P_{75}) is referred to as $Q3$ or the third quartile because three-quarters of the scores fall below this point.

Percentiles are easy to compute and intuitively appealing to laypersons and professionals alike. It is not surprising, then, that percentiles are the most common type of raw score transformation encountered in psychological testing. Almost any kind of test result can be reported as a percentile, even when other transformations are the primary goal of testing. For example, intelligence tests are used to obtain IQ scores—a kind of transformation discussed subsequently —but also yield percentile scores, too. Thus, an IQ of 130 corresponds to a percentile of 98, meaning that the score is not only well above average but, more precisely, also exceeds 98 percent of the standardization sample.

Percentile scores do have one major drawback: They distort the underlying measurement scale, especially at the extremes. A specific example will serve to clarify this point. Consider a hypothetical instance in which four persons obtain the following percentiles on a test: 50, 59, 90, and 99. (Remember that we are speaking here of percentiles, not percent correct.) The first two persons differ by 9 percentile points (50 versus 59) and so do the last two persons (90 versus 99). The untrained observer might assume, falsely, that the first two persons differed in underlying raw score points by the same amount as the last two persons. An inspection of **Figure 3.5** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03fig5) reveals the fallacy of this assumption. The difference in underlying raw score points between percentiles of 90 and 99 is far greater than between percentiles of 50 and 59.

Standard Scores

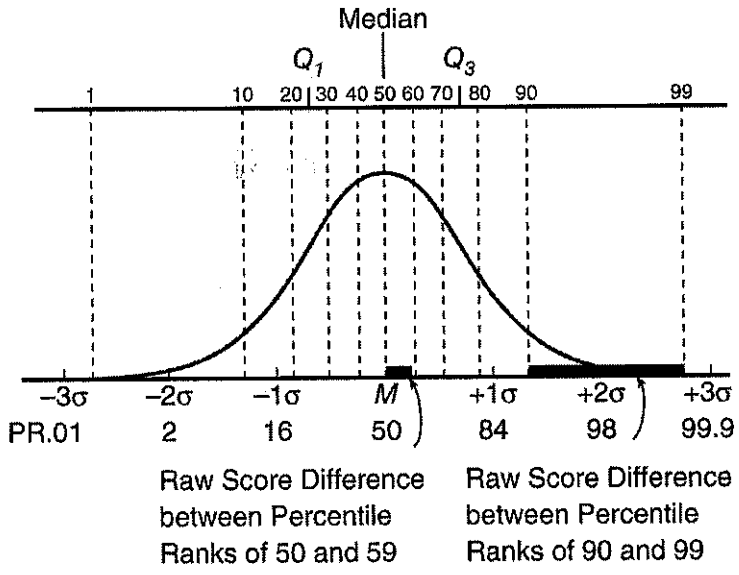


FIGURE 3.5 Percentile Ranks in a Normal Distribution

Although percentiles are the most popular type of transformed score, standard scores exemplify the most desirable psychometric properties. A standard score uses the standard deviation of the total distribution of raw scores as the fundamental unit of measurement. The **standard score** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss308) expresses the distance from the mean in standard deviation units. For example, a raw score that is exactly one standard deviation above the mean converts to a standard score of +1.00. A raw score that is exactly one-half a standard deviation below the mean converts to a standard score of −.50. Thus, a standard score not only expresses the magnitude of deviation from the mean, but the direction of departure (positive or negative) as well.

Computation of an examinee's standard score (also called a *z* score) is simple: Subtract the mean of the normative group from the examinee's raw score and then divide this difference by the standard deviation of the normative group. **Table 3.3** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03tab3)** illustrates the computation of *z* scores for three subjects of widely varying ability on a hypothetical test.

Standard scores possess the desirable psychometric property of retaining the relative magnitudes of distances between successive values found in the original raw scores. This is because the distribution of standard scores has exactly the same shape as the distribution of raw scores. As a consequence, the use of standard scores does not distort the underlying measurement scale. This fidelity of the transformed measurement scale is a major advantage of standard scores over percentiles and percentile ranks. As previously noted, percentile scores are very distorting, especially at the extremes.

TABLE 3.3 Computation of Standard Scores on a Hypothetical Test

For the normative sample: $M = 50, SD = 8$

$$\text{Standard score} = z = \frac{X - M}{SD}$$

Person A: raw score of 35 (below average)

$$z = \frac{35 - 50}{8} = -1.88$$

Person B: raw score of 50 (exactly average)

$$z = \frac{50 - 50}{8} = 0.00$$

Person C: raw score of 70 (above average)

$$z = \frac{70 - 50}{8} = +2.50$$

A specific example will serve to illustrate the nondistorting feature of standard scores. Consider four raw scores of 55, 60, 70, and 80 on a test with mean of 50 and standard deviation of 10. The first two scores differ by 5 raw score points, while the last two scores differ by 10 raw score points—twice the difference of the first pair. When the raw scores are converted to standard scores, the results are +0.50, +1.00, +2.00, and +3.00, respectively. The reader will notice that the first two scores differ by 0.50 standard scores, while the last two scores differ by 1.00 standard scores—twice the difference of the first pair. Thus, standard scores always retain the relative magnitude of differences found in the original raw scores.

Standard score distributions possess important mathematical properties that do not exist in the raw score distributions. When each of the raw scores in a distribution is transformed to a standard score, the resulting collection of standard scores always has a mean of zero and a variance of 1.00. Because the standard deviation is the square root of the variance, the standard deviation of standard scores ($\sqrt{1.00}$) is necessarily 1.00 as well.

One reason for transforming raw scores into standard scores is to depict results on different tests according to a common scale. If two distributions of test scores possess the same form, we can make direct comparisons on raw scores by transforming them to standard scores. Suppose, for example, that a first-year college student earned 125 raw score points on a spatial thinking test for which the normative sample averaged 100 points (with SD of 15 points). Suppose, in addition, he earned 110 raw score points on a vocabulary test for which the normative sample averaged 90 points (with SD of 20 points). In which skill area does he show greater aptitude, spatial thinking or vocabulary?

If the normative samples for both tests produced test score distributions of the same form, we can compare spatial thinking and vocabulary scores by converting each to standard scores. The spatial thinking standard score for our student is (125 – 100)/15 or +1.67, whereas his vocabulary standard score is (110 – 90)/20 or +1.00. Relative to the normative samples, the student has greater aptitude for spatial thinking than vocabulary.

But a word of caution is appropriate when comparing standard scores from different distributions. If the distributions do not have the same form, standard score comparisons can be very misleading. We illustrate this point with **Figure 3.6** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03fig6)** , which depicts two distributions: one markedly skewed with average score of 30 (SD of 10) and another normally distributed with average score of 60 (SD of 8). A raw score of 40 on the first test and a raw score of 68 on the second test both translate to identical standard scores of +1.00. Yet, a standard score of 1.00 on the first test exceeds 92 percent of the normative sample, while the equivalent standard score on the second test exceeds only 84 percent of the normative sample. When two distributions of test scores do not possess the same form, equivalent standard scores do not signify comparable positions within the respective normative samples.

T Scores and Other Standardized Scores

Many psychologists and educators appreciate the psychometric properties of standard scores but regard the decimal fractions and positive/negative signs (e.g., $z = -2.32$) as unnecessary distractions. In response to these concerns, test specialists have devised a number of variations on standard scores that are collectively referred to as *standardized scores.*

From a conceptual standpoint, standardized scores are identical to standard scores. Both kinds of scores contain exactly the same information. The shape of the distribution of scores is not affected, and a plot of the relationship between standard and standardized scores is always a straight line. However, standardized scores are always expressed as positive whole numbers (no decimal fractions or negative signs), so many test users prefer to depict test results in this form.

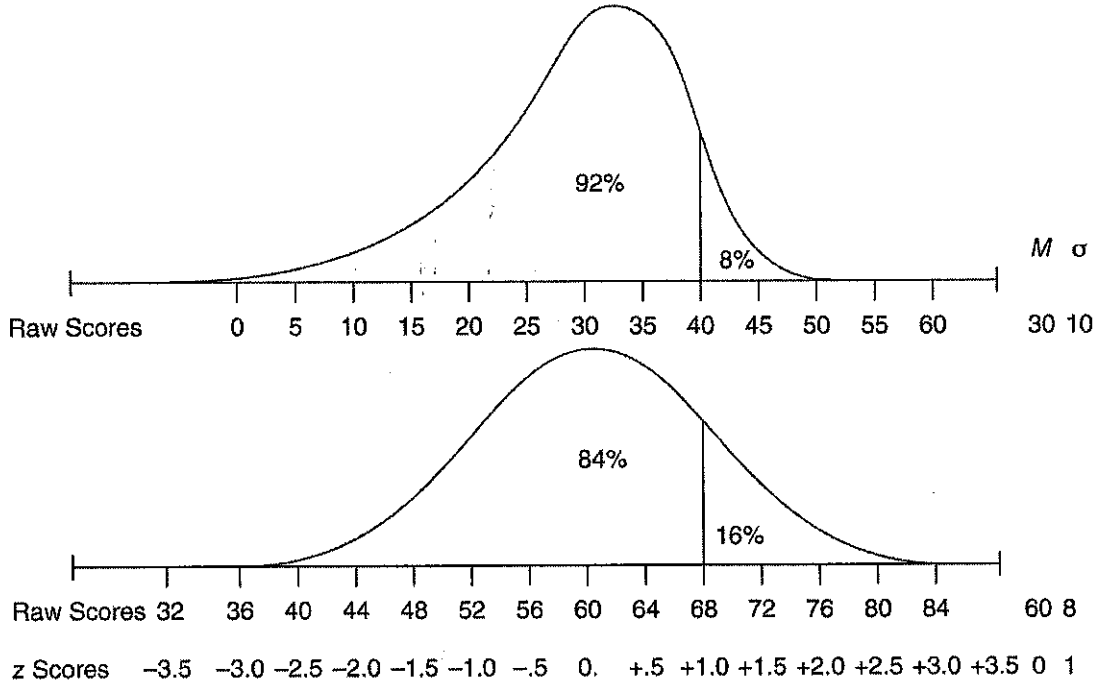


FIGURE 3.6 Relationships between Raw Scores, z Scores, and Relative Standing for Two Distributions of Markedly Different Form

Standardized scores eliminate fractions and negative signs by producing values other than zero for the mean and 1.00 for the standard deviation of the transformed scores. The mean of the transformed scores can be set at any convenient value, such as 100 or 500, and the standard deviation at, say, 15 or 100. The important point about standardized scores is that we can transform any distribution to a preferred scale with predetermined mean and standard deviation.

One popular kind of standardized score is the ***T* score** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss337), which has a mean of 50 and a standard deviation of 10. *T* score scales are especially common with personality tests. For example, on the MMPI, each clinical scale (e.g., Depression, Paranoia) is converted to a common metric for which 50 is the average score and 10 is the standard deviation for the normative sample.

To transform raw scores to *T* scores, we use the following formula:

$$T = \frac{10(X - M)}{\text{SD}} + 50$$

The term $(X - M)/\text{SD}$ is, of course, equivalent to *z*, so we can rewrite the formula for *T* as a simple transformation of *z*:

$$T = 10z + 50$$

For any distribution of raw scores, the corresponding *T* scores will have an average of 50. In addition, for most distributions the vast majority of *T* scores will fall between values of 20 and 80, that is, within three standard deviations of the mean. Of course, *T* scores outside this range are entirely possible and perhaps even likely in special populations. In clinical settings it is not unusual to observe very high *T* scores—even as high as 90—on personality inventories such as the MMPI.

Standardized scores can be tailored to produce any mean and standard deviation. However, to eliminate negative standardized scores, the preselected mean should be at least five times as large as the standard deviation. In practice, test developers rely upon a few preferred values for means and standard deviations of standardized scores, as outlined in **Table 3.4** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03tab4).

Normalizing Standard Scores

As previously noted, psychologists and educators prefer to deal with normal distributions because the statistical properties of the normal curve are well known and standard scores from these distributions can be directly compared. Perhaps the reader has wondered what recourse is available to test developers who find that their tests produce an asymmetrical distribution of scores in the normative sample. Fortunately, distributions of scores that are skewed or otherwise nonnormal can be transformed or normalized to fit a normal curve. Although test specialists have devised several methods for transmuting a nonnormal distribution into a normal one, we will discuss only the most popular approach—the conversion of percentiles to normalized standard scores. Oddly enough, it is easier to explain this approach if we first describe the reverse process: conversion of standard scores to percentiles.

TABLE 3.4 Means and Standard Deviations of Common Standardized Scores

Type of Measure	*Specific Examples*	*Mean*	*Standard Deviation*
Full Scale	IQ WAIS-IV	100	15
IQ Test Subscales	Vocabulary, Block Design	10	3
Personality Test Scales	MMPI-2 Depression, Paranoia	50	10
Aptitude Tests	Graduate Record Exam, Scholastic Assessment Tests	500	100

We have noted that a normal distribution of raw scores has, by definition, a distinct, mathematically defined shape (**Figure 3.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig3)). In addition, we have pointed out that transforming a group of raw scores to standard scores leaves the original form of a distribution unchanged. Thus, if a collection of raw scores is normally distributed, the resulting standard scores will obey the normal curve, too.

We also know that the mathematical properties of the normal distribution are precisely calculable. Without going into the details of computation, it should be obvious that we can determine the percentage of cases falling below any particular standard score. For example, in **Figure 3.3** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec2#ch03fig3) , a standard score of −2.00 (designated as -2σ) exceeds 2.14 percent of the cases. Thus, a standard score of −2.00 corresponds to a percentile of 2.14. In like manner, any conceivable standard score can be expressed in terms of its corresponding percentile. Appendix D lists percentiles for standard scores and several other transformed scores.

Producing a **normalized standard score** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss222) is accomplished by working in the other direction. Namely, we use the percentile for each raw score to determine its corresponding standard score. If we do this for each and every case in a nonnormal distribution, the resulting distribution of standard scores will be normally distributed. Notice that in such a normalized standard score distribution, the standard scores are not calculated directly from the usual computational formula but are determined indirectly by first computing the percentile and then ascertaining the equivalent standard score.

The conversion of percentiles to normalized standard scores might seem an ideal solution to the problem of unruly test data. However, there is a potentially serious drawback: Normalized standard scores are a nonlinear transformation of the raw scores. Thus, mathematical relationships established with the raw scores may not hold true for the normalized standard scores. In a markedly skewed distribution, it is even possible that a raw score that is significantly below the mean might conceivably have a normalized standard score that is above the mean.

In practice, normalized standard scores are used sparingly. Such transformations are appropriate only when the normative sample is large and representative and the raw score distribution is only mildly nonnormal. Incidentally, the most likely cause of these nonnormal score distributions is inappropriate difficulty level in the test items, such as too many difficult or easy items.

There is a catch-22 here, in that mildly non-normal distributions are not changed much when they are normalized, so little is gained in the process. Ironically, normalized standard scores produce the greatest change with markedly nonnormal distributions. However, when the raw score distribution is markedly nonnormal, test developers are better advised to go back to the drawing board and adjust the difficulty level of test items so as to produce a normal distribution, rather than succumb to the partial statistical fix of normalized standard scores.

Stanines, Stens, and C Scale

Finally, we give brief mention to three raw score transformations that are mainly of historical interest. The stanine (standard nine) scale was developed by the United States Air Force during World War II. In a **stanine scale** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss312) , all raw scores are converted to a single-digit system of scores ranging from 1 to 9. The mean of stanine scores is always 5, and the standard deviation is approximately 2. The transformation from raw scores to stanines is simple: The scores are ranked from lowest to highest, and the bottom 4 percent of scores convert to a stanine of 1, the next 7 percent convert to a stanine of 2, and so on (see **Table 3.5** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03tab5)). The main advantage of stanines is that they are restricted to single-digit numbers. This was a considerable asset in the premodern computer era in which data was keypunched on Hollerith cards that had to be physically carried and stored on shelves. Because a stanine could be keypunched in a single column, far fewer cards were required than if the original raw scores were entered.

TABLE 3.5 Distribution Percentages for Use in Stanine Conversion

Percentage	4	7	12	17	20	17	12	7	4
Stanine	1	2	3	4	5	6	7	8	9

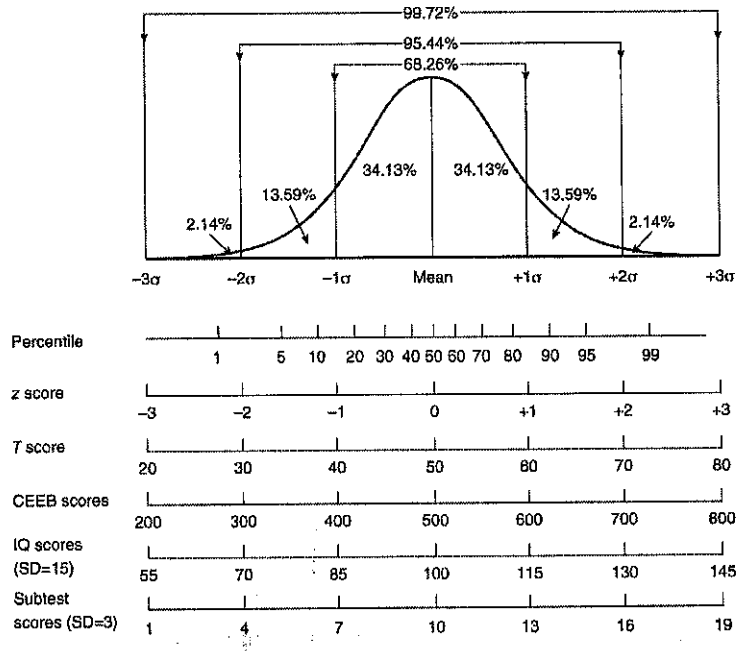


FIGURE 3.7 Equivalencies between Common Raw Score Transformation in a Normal Distribution

Statisticians have proposed several variations on the stanine theme. Canfield (**1951** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib266)**) proposed the 10-unit **sten scale** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss314)** , with 5 units above and 5 units below the mean. Guilford and Fruchter (**1978** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib668)**) proposed the **C scale** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss38)** consisting of 11 units. Although stanines are still in widespread use, variants such as the sten and C scale never roused much interest among test developers.

A Summary of Statistically Based Norms

We have alluded several times to the ease with which standard scores, *T* scores, stanines, and percentiles can be transformed into each other, especially if the underlying distribution of raw scores is normally distributed. In fact, the exact form in which scores are reported is largely a matter of convention and personal preference. For example, a WAIS-III IQ of 115 could also be reported as a standard score of +1.00, or a *T* score of 60, or a percentile rank of 84. All of these results convey exactly the same information.[1] **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03fn01)** **Figure 3.7** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec3#ch03fig7)** summarizes the relationships that exist between the most commonly used statistically based norms.

This ends the brief introduction to the many techniques by which test data from a normative sample can be statistically summarized and transformed. We should never lose sight of the overriding purpose of these statistical transmutations, namely, to help the test user make sense out of one individual's score in relation to an appropriate comparison group.

But what is an appropriate comparison group? What characteristics should we require in our norm group subjects? How should we go about choosing these subjects? How many subjects do we need? These are important questions that influence the relevance of test results just as much as proper item selection and standardized testing procedure. In the remainder of this topic, we examine the procedures involved in selecting a norm group.

[1] A WAIS-III IQ of 115 also can be expressed as a stanine of 7. However, it is worth noting that some information is lost when scores are reported as stanines. Note that IQs in the range of 111 to 119 *all* convert to a stanine of 7. Thus, if we are told only that an individual has achieved at the 7th stanine on an intelligence test, we do not know the exact IQ equivalent.

3.4 SELECTING A NORM GROUP

When choosing a norm group, test developers strive to obtain a representative cross-section of the population for whom the test is designed (**Petersen, Kolen, & Hoover, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1286)). In theory, obtaining a representative norm group is straightforward and simple. Consider a scholastic achievement test designed for sixth graders in the United States. The relevant population is all sixth graders coast to coast and in Alaska and Hawaii. A representative cross-section of these potential subjects could be obtained by computerized random sampling of 10,000 or so of the millions of eligible children. Each child would have an equal chance of being chosen to take the test; that is, the selection strategy would be simple **random sampling** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss265) . The results for such a sample would comprise an ideal source of normative data. With a large random sample, it is almost certain that the diversities of ethnic background, social class, geographic location, urban versus rural setting, and so on would be proportionately represented in the sample.

In the real world, obtaining norm samples is never as simple and definitive as the hypothetical case previously outlined. Researchers do not have a complete list of every sixth grader in the nation, and even if they did, test developers could not compel every randomly selected child to participate in the standardization of a test. Questions of cost arise, too. Psychometricians must be paid to administer the tests to the norm group. Test developers may opt for a few hundred representative subjects instead of a larger number.

To help ensure that smaller norm groups are truly representative of the population for which the test was designed, test developers employ **stratified random sampling** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss316) . This approach consists of stratifying, or classifying, the target population on important background variables (e.g., age, sex, race, social class, educational level) and then selecting an appropriate percentage of persons at random from each stratum. For example, if 12 percent of the relevant population is African American, then the test developer chooses subjects randomly but with the constraint that 12 percent of the norm group is also African American.

In practice, very few test developers fully emulate either random sampling or stratified random sampling in the process of selecting the norm group. What is more typical is a good faith effort to pick a diverse and representative sample from strong and weak schools, minority and white neighborhoods, large and small cities, and northern, eastern, central, and southern communities. If this sample then embodies about the same percentage of minorities, city dwellers, and upper- and lower-class families as the national census, then the test developer feels secure that the norm group is representative.

There is an important lesson in the uncertainties, compromises, and pragmatics of norm group selection; namely, psychological test norms are not absolute, universal, or timeless. They are relative to one historical era and the particular normative population from which they were derived. We will illustrate the ephemeral nature of normative statistics in a later section when we show how a major IQ test normed at a national average of 100 in 1974 yielded a national average of 107 in 1988. Even norms that are selected with great care and based on large samples can become obsolete in a decade—sometimes less.

Age and Grade Norms

As we grow older, we change in measurable ways, for better or worse. This is obviously true in childhood, when intellectual skills improve visibly from one month to the next. In adulthood, personal change is slower but still discernible. We expect, for example, that adults will show a more mature level of vocabulary with each passing decade (**Gregory & Gernert, 1990** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib652)).

An **age norm** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss05) depicts the level of test performance for each separate age group in the normative sample. The purpose of age norms is to facilitate same-aged comparisons. With age norms, the performance of an examinee is interpreted in relation to standardization subjects of the same age. The age span for a normative age group can vary from a month to a decade or more, depending on the degree to which test performance is age-dependent. For characteristics that change quickly with age—such as intellectual abilities in childhood—test developers might report separate test norms for narrowly defined age brackets, such as four-month intervals. This allows the examiner, for example, to compare test results of a child who is 5 years and 2 months old (age 5-2) to the normative sample of children ranging from age 5–0 to age 5–4. By contrast, adult characteristics change more slowly and it might be sufficient to report normative data by 5- or 10-year age intervals.

Grade norms are conceptually similar to age norms. A **grade norm** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss136) depicts the level of test performance for each separate grade in the normative sample. Grade norms are rarely used with ability tests. However, these norms are especially useful in school settings when reporting the achievement levels of schoolchildren. Since academic achievement in many content areas is heavily dependent on grade-based curricular exposure, comparing a student against a normative sample from the same grade is more appropriate than using an age-based comparison.

Local and Subgroup Norms

With many applications, local or subgroup norms are needed to suit the specific purpose of a test. **Local norms** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss189) are derived from representative local examinees, as opposed to a national sample. Likewise, **subgroup norms** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss317) consist of the scores obtained from an identified subgroup (African Americans, Hispanics, females), as opposed to a diversified national sample. As an example of local norms in action, the admissions officer of a junior college that attracts mainly local residents might prefer to consult statewide norms rather than national norms on a scholastic achievement test.

As a general rule, whenever an identifiable subgroup performs appreciably better or worse on a test than the more broadly defined standardization sample, it may be helpful to construct supplementary subgroup norms. The subgroups can be formed with respect to sex, ethnic background, geographical region, urban versus rural environment, socio economic level, and many other factors.

Whether local or subgroup norms are beneficial depends on the purpose of testing. For example, ethnic norms for standardized intelligence tests may be superior to nationally based norms in predicting competence within the child's nonschool environment. However, ethnic norms may not predict how well a child will succeed in mainstream public school instructional programs (**Mercer & Lewis, 1978** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1131)). Thus, local and subgroup norms must be used cautiously.

Expectancy Tables

One practical form that norms may take is an expectancy table. An **expectancy table** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss104) portrays the established relationship between test scores and expected outcome on a relevant task (**Harmon, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib705)). Expectancy tables are

especially useful with predictor tests used to forecast well-defined criteria. For example, an expectancy table could depict the relationship between scores on a scholastic aptitude test (predictor) and subsequent college grade point average (criterion).

Expectancy tables are always based on the previous predictor and criterion results for large samples of examinees. The practical value of tabulating normative information in this manner is that new examinees receive a probabilistic preview of how well they are likely to do on the criterion. For example, high school examinees who take a scholastic aptitude test can be told the statistical odds of achieving a particular college grade point average.

Based on 7,835 previous examinees who subsequently attended a major university, the expectancy table in **Table 3.6 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec4#ch03tab6)** provides the probability of achieving certain first-year college grades as a function of score on the American College Testing (ACT) examination. The ACT test is typically given to high school seniors who have expressed an interest in attending college. The first column of the table shows ACT test scores, divided into 10 class intervals. The second column gives the number of students whose scores fell into each interval. The remaining entries in each row show the percentage of students within each test-score interval who subsequently received college grade points within a designated range. For example, of the 117 students who scored 31 to 33 points on the ACT, only 2 percent received a first-year college grade point average below 1.50, while 64 percent earned superlative grades of 3.50 up to a perfect A or 4.00. At the other extreme, of the 102 students who scored below 10 points on the ACT, fully 80 percent (60 percent plus 20 percent) received first-year college grades below a C average of 2.00.

TABLE 3.6 Expectancy Table Showing Relation between ACT Composite Scores and First-Year College Grades for 7,835 Students at a Major State University

		Grade Point Average (4.00 Scale)					
ACT Test Score	*Number of Cases*	*0.00–1.49*	*1.50–1.99*	*2.00–2.49*	*2.50–2.99*	*3.00–3.49*	*3.50–4.00*
34–36	3	0	0	33	0	0	67
31–33	117	2	2	4	9	19	64
28–30	646	10	6	10	17	23	35
25–27	1,458	12	10	16	19	24	19
22–24	1,676	17	10	22	20	20	11
19–21	1,638	23	14	25	18	16	4
16–18	1,173	31	17	24	15	11	3
13–15	690	38	18	25	12	6	1
10–12	332	54	16	20	6	3	1
below 10	102	60	20	13	8	0	0

Note: Some rows total to more than 100 percent because of rounding errors.

Source: Courtesy of Archie George, Management Information Services, University of Idaho.

Of course, expectancy tables do not foreordain how new examinees will do on the criterion. In an individual case, it is conceivable that a low–ACT scoring student might beat the odds and earn a 4.00 college grade point average. More commonly, though, new examinees discover that expectancy tables provide a broadly accurate preview of criterion performance.

But there are some exceptional instances in which expectancy tables can become inaccurate. An expectancy table is always based on the previous performance of a large and representative sample of examinees whose test performances and criterion outcomes reflected existing social conditions and institutional policies. If conditions or policies change, an expectancy table can become obsolete and misleading.

3.5 CRITERION-REFERENCED TESTS

We close this unit with a brief mention of an alternative to norm-referenced tests, namely, criterion-referenced tests. These two kinds of tests differ in their intended purposes, the manner in which content is chosen, and the process of interpreting results (**Hambleton & Zenitsky, 2003** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib689) ; **Bond, 1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib176) ; **Frechtling, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib537) ; **Popham, 1978** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1315)).

The purpose of a norm-referenced test is to classify examinees, from low to high, across a continuum of ability or achievement. Thus, a norm-referenced test uses a representative sample of individuals—the norm group or standardization sample—as its interpretive framework. Examiners might want to classify individuals in this way for purposes of selection to a specialized curriculum or placement in remedial or gifted programs. In a classroom setting, a teacher might use a norm-referenced test to assign students to different reading levels or math instructional groups (**Bond, 1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib176)).

Whereas norm-referenced tests are used to rank students along a continuum in comparison to one another, criterion-referenced tests are used to compare examinees' accomplishments to a predefined performance standard. For example, consider a hypothetical school system in which fourth graders are expected to master the addition of pairs of two-digit numbers (e.g., 23 + 19 = 42). Perhaps the performance standard is set at 80 percent accuracy when doing 10 such addition problems in a 15-minute time period. Results for a specific fourth grader are then descriptively stated as a particular percentage (e.g., 70 percent). While it is possible to compare this result to the predetermined *standard*, no comparison is made to other *students*. In fact, it is entirely possible (and even desirable) for all students to exceed the standard.

Criterion-referenced tests represent a fundamental shift in perspective. The focus is on what the test taker can do rather than on comparisons to the performance levels of others. Thus, criterion-referenced tests identify an examinee's relative mastery (or nonmastery) of specific, predetermined competencies. These kinds of tests are increasingly popular in educational systems, where they are used to evaluate how well students have mastered the academic skills expected at each grade level. This information, in turn, provides a basis for intervention with students who are lagging behind. In addition, system-wide results of criterion-referenced tests can be used to evaluate the curriculum and to determine how well individual schools are teaching the curriculum.

A major difference between norm-referenced tests and criterion-referenced tests is the manner in which test content is chosen. In a norm-referenced test, items are chosen so that they provide maximal discrimination among respondents along the dimension being measured. Within this framework, well-defined psychometric principles are used to identify ideal items according to difficulty level, correlation with the total score, and other properties. In contrast, with a criterion-referenced test, the content is selected on the basis of its relevance in the curriculum. This involves the judgment and consensus of educators and other stakeholders in the educational enterprise. In **Table 3.7** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec5#ch03tab7) , we have summarized and compared some distinctive characteristics of criterion-referenced and norm-referenced tests.

Criterion-referenced tests are best suited to the testing of basic academic skills (e.g., reading level, computation skill) in educational settings. However, these kinds of instruments are largely inappropriate for testing higher-level abilities because it is difficult to formulate specific objectives for such content domains. Consider a particular case: How could we develop a criterion-referenced test for expert computer programming? It would be difficult to propose specific behaviors that all expert computer programmers would possess and, therefore, nearly impossible to construct a criterion-referenced test for this high-level skill. Berk (**1984** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib148)) discusses the technical problems in the construction and evaluation of criterion-referenced tests.

TABLE 3.7 Distinctive Characteristics of Criterion-Referenced and Norm-Referenced Tests

Dimension	*Criterion-Referenced Tests*	*Norm-Referenced Tests*
Purpose	Compare examinees' performance to a standard	Compare examinees' performance to one another
Item Content	Narrow domain of skills with real-world relevance	Broad domain of skills with indirect relevance
Item Selection	Most items of similar difficulty level	Items vary widely in difficulty level
Interpretation of Scores	Scores usually expressed as a percentage, with passing level predetermined	Scores usually expressed as a standard score, percentile, or grade equivalent

A common application of criterion-referenced tests (CRTs) is in educational settings where they are used to determine whether students have met the minimum or basic standards in curriculum areas such as algebra, reading, or science. As noted, students are compared to a standard, not to one another. CRTs allow for the possibility that everyone might pass. At first glance, they might appear to be more equitable than norm-referenced tests which feature comparisons among students. However, as noted by FairTest, the National Center for Open and Fair Testing (**www.fairtest.org** (http://www.fairtest.org)), whether CRTs are really fair depends upon how the cut-off scores are determined:

> On a standardized CRT (one taken by students in many schools), the passing or "cut-off" score is usually set by a committee of experts, while in a classroom the teacher sets the passing score. In both cases, deciding the passing score is subjective, not objective. Sometimes cut scores have been set in a way that maximizes the number of low-income or minority students who fail the test. A small change in the cut score would not change the meaning of the test but would greatly increase minority pass rates. (**www.fairtest.org** (http://www.fairtest.org))

Criterion-referenced tests can be used for specific classroom objectives (e.g., meeting a minimal level of proficiency in spelling for sixth graders) or for more far-reaching, high-stakes purposes such as determining graduation from high school. An example of the latter is the AIMS Test (Arizona Instrument to Measure Standards), used statewide in Arizona as a high school exit exam (**Arizona Senate Research Staff, 2008** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib60)). The test is designed to measure academic achievement in reading, writing, and math. The *minimum* passing level is mastery at a 10th grade level in all subjects for graduating seniors. Exemptions are granted for some students in special education.

Ultimately, individual Arizona public schools are beholden to these criterion-referenced standards as well. The AIMS Test is given in grades 3 through 8 and also serves as the benchmark for graduation in the senior year. The state legislation authorizing AIMS also stipulates that a school is making adequate yearly progress if at least 90 percent of its students pass the AIMS test at their grade level, or if the percentage passing is higher than the previous year (**Arizona**

Senate Research Staff, 2008 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib60)). Based on these data, schools receive a label of either (1) excelling; (2) highly performing; (3) performing; (4) underperforming; or (5) failing. Underperforming and failing schools face outside review. Certainly the AIMS Test is an example of high-stakes testing, as discussed in the first chapter.

Another concern is the degree to which the test matches the curriculum. Many state tests are developed by a committee of experts who have only general ideas about what students might be taught. The tests that emerge from the committee might not match the curricula for specific school systems. Thus, they might include areas that some students have not studied.

TOPIC 3B Concepts of Reliability

3.6 Classical Test Theory and the Sources of Measurement Error
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec6#ch03lev1sec6)

3.7 Sources of Measurement Error (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec7#ch03lev1sec7)

3.8 Measurement Error and Reliability (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec8#ch03lev1sec8)

3.9 The Reliability Coefficient (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec9#ch03lev1sec9)

3.10 The Correlation Coefficient (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec10#ch03lev1sec10)

3.11 The Correlation Coefficient as a Reliability Coefficient
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec11#ch03lev1sec11)

3.12 Reliability as Temporal Stability (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec12#ch03lev1sec12)

3.13 Reliability as Internal Consistency (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec13#ch03lev1sec13)

3.14 Item Response Theory (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03lev1sec14)

3.15 The New Rules of Measurement (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec15#ch03lev1sec15)

3.16 Special Circumstances in the Estimation of Reliability
(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec16#ch03lev1sec16)

3.17 The Interpretation of Reliability Coefficients (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec17#ch03lev1sec17)

3.18 Reliability and the Standard Error of Measurement (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec18#ch03lev1sec18)

Reliability refers to the attribute of consistency in measurement. However, reliability is seldom an all-or-none matter; more commonly it is a question of degree. Very few measures of physical or psychological characteristics are completely consistent, even from one moment to the next. For example, a person who steps on a scale twice in quick succession might register a weight of 145½ pounds the first time and 145¾ pounds the second. The same individual might take two presumably equivalent forms of an IQ test and score 114 on one and 119 on the other. Two successive measures of speed of response—pressing a key quickly whenever the letter *X* appears on a microcomputer screen—might produce a reaction time of 223 milliseconds on the first trial and 341 milliseconds on the next. We see in these examples a pattern of consistency—the pairs of measurements are not completely random—but different amounts of inconsistency are evident, too. In the short run, measures of weight are highly consistent, intellectual test scores are moderately stable, but simple reaction time is somewhat erratic.

The concept of **reliability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss274) is best viewed as a continuum ranging from minimal consistency of measurement (e.g., simple reaction time) to near-perfect repeatability of results (e.g., weight). Most psychological tests fall somewhere in between these two extremes. With regard to tests, an acceptable degree of reliability is more than an academic matter. After all, it would be foolish and unethical to base important decisions on test results that are not repeatable.

Psychometricians have devised several statistical methods for estimating the degree of reliability of measurements, and we will explore the computation of such reliability coefficients in some detail. But first we examine a more fundamental issue to help clarify the meaning of reliability: What are the sources of consistency and inconsistency in psychological test results?

3.6 CLASSICAL TEST THEORY AND THE SOURCES OF MEASUREMENT ERROR

The theory of measurement introduced here has been called the classical test theory because it was developed from simple assumptions made by test theorists since the inception of testing. This approach is also called the *theory of true and error scores*, for reasons explained below. Charles Spearman (**1904** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1526)) laid down the foundation for the theory that was subsequently extended and revised by contemporary psychologists (**Feldt & Brennan, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib495); **Lord & Novick, 1968** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1006); **Kline, 1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib900)). We should mention that a rival model does exist and is slowly supplanting classical test theory as a basis for test development. Item response theory, or latent trait theory (Embretson & Hershberger, 1999), is an appealing alternative to classical test theory. We close this chapter with a brief review of item response theory. However, classical test theory was the basis for test development throughout most of the twentieth century. Accordingly, we begin our coverage with this model.

The basic starting point of the **classical theory of measurement** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss46) is the idea that test scores result from the influence of two factors:

- Factors that contribute to consistency. These consist entirely of the stable attributes of the individual, which the examiner is trying to measure.
- Factors that contribute to inconsistency. These include characteristics of the individual, test, or situation that have nothing to do with the attribute being measured, but that nonetheless affect test scores.

It should be clear to the reader that the first factor is desirable because it represents the true amount of the attribute in question, while the second factor represents the unavoidable nuisance of error factors that contribute to inaccuracies of measurement. We can express this conceptual breakdown as a simple equation:

$$X = T + e$$

where X is the obtained score, T is the **true score** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss336), and e represents errors of measurement.

Errors in measurement, thus, represent discrepancies between the obtained scores and the corresponding true scores:

$$e = X - T$$

Notice in the preceding equations that errors of measurement e can be either positive or negative. If e is positive, the obtained score X will be higher than the true score T. Conversely, if e is negative, the obtained score will be lower than the true score. Although it is impossible to eliminate all **measurement error** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss193), test developers do strive to minimize this psychometric nuisance through careful attention to the sources of measurement error outlined in the following section.

Finally, it is important to stress that the true score is never known. As the reader will discover, we can obtain a probability that the true score resides within a certain interval and we can also derive a best estimate of the true score. However, we can never know the value of a true score with certainty.

3.7 SOURCES OF MEASUREMENT ERROR

As indicated by the formula $X = T + e$, measurement error e is everything other than the true score that makes up the obtained test score. Errors of measurement can arise from innumerable sources (**Feldt & Brennan, 1989** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib495)**). Stanley (**1971** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1545)**) provides an unusually thorough list. We will outline only the most important and likely contributions here: item selection, test administration, test scoring, and systematic errors of measurement.

Item Selection

One source of measurement error is the instrument itself. A test developer must settle on a finite number of items from a potentially infinite pool of test questions. Which questions should be included? How should they be worded? Item selection is crucial to the accuracy of measurement.

Although psychometricians strive to obtain representative test items, the particular set of questions chosen for a test might not be equally fair to all persons. A hypothetical and deliberately extreme example will serve to illustrate this point: Even a well-prepared student might flunk a classroom test that emphasized the obscure footnotes in the textbook. By contrast, an ill-prepared but curious student who studied only the footnotes might do very well on such an exam. The scores for both persons would reflect massive amounts of measurement error. Remember in this context that the true score is what the student really knows. For the conscientious student, the obtained score would be far lower than the true score because of a hefty dose of negative measurement error. For the serendipitous second student, the obtained score would be far higher than the true score, owing to the positive measurement error.

Of course, in a well-designed test the measurement error from item sampling will be minimal. However, a test is always a sample and never the totality of a person's knowledge or behavior. As a result, item selection is always a source of measurement error in psychological testing. The best a psychometrician can do is minimize this unwanted nuisance by attending carefully to issues of test construction. We discuss technical aspects of item selection in **Topic 4B** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch04lev1sec6#ch04box2)** , Test Construction.

Test Administration

Although examiners usually provide an optimal and standardized testing environment, numerous sources of measurement error may nonetheless arise from the circumstances of administration. Examples of general environmental conditions that may exert an untoward influence on the accuracy of measurement include uncomfortable room temperature, dim lighting, and excessive noise. In some cases it is not possible to anticipate the qualities of the testing situation that will contribute to measurement error. Consider this example: An otherwise lackluster undergraduate correctly answers a not very challenging information item, namely, "Who wrote *Canterbury Tales*?" When queried later whether he had read any Chaucer, the student replies, "No, but you've got that book right behind you on your bookshelf."

Momentary fluctuations in anxiety, motivation, attention, and fatigue level of the test taker may also introduce sources of measurement error. For example, an examinee who did not sleep well the night before might lack concentration and, therefore, misread questions. A student distracted by temporary emotional distress might inadvertently respond in the wrong columns of the answer sheet. The classic nightmare in this regard is the test taker who skips a question—let us say, question number 19—but forgets to leave the corresponding part of the answer sheet blank. As a result, all the subsequent answers are off by one, with the response to question 20 entered on the answer sheet as item 19, and so on.

The examiner, too, may contribute to measurement error in the process of test administration. In an orally administered test, an unconscious nod of the head by the tester might convey that the examinee is on the right track, thereby guiding the test taker to the correct response. Conversely, a terse and abrupt examiner may intimidate a test taker who would otherwise volunteer a correct answer.

Test Scoring

Whenever a psychological test uses a format other than machine-scored multiple-choice items, some degree of judgment is required to assign points to answers. Fortunately, most tests have well-defined criteria for answers to each question. These guidelines help minimize the impact of subjective judgment in scoring (**Gregory, 1987** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib645)**). However, subjectivity of scoring as a source of measurement error can be a serious problem in the evaluation of projective tests or essay questions. With regard to projective tests, Nunnally (**1978** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1243)**) points out that the projective tester might undergo an evolutionary change in scoring criteria over time, coming to regard a particular type of response as more and more pathological with each encounter.

Systematic Measurement Error

The sources of inaccuracy previously discussed are collectively referred to as **unsystematic measurement error**, meaning that their effects are unpredictable and inconsistent. However, there is another type of measurement error that constitutes a veritable ghost in the psychometric machine. A **systematic measurement error** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss321)** arises when, unknown to the test developer, a test consistently measures something other than the trait for which it was intended. Systematic measurement error actually is a problem for test validity, as discussed in the next chapter. Yet, we mention it here because it does contribute to inaccuracies of measurement.

Suppose, for example, that a scale to measure social introversion also inadvertently taps anxiety in a consistent fashion. In this case, the equation depicting the relationship between observed scores, true scores, and sources of measurement error would be

$$X = T + e_s + e_u$$

where X is the obtained score, T is the true score, e_s is the systematic error due to the anxiety subcomponent, and e_u is the collective effect of the unsystematic measurement errors previously outlined.

Because by definition their presence is initially undetected, systematic measurement errors may constitute a significant problem in the development of psychological tests. However, if psychometricians use proper test development procedures discussed in **Topic 4B** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch04lev1sec6#ch04box2)** , Test Construction, the impact of systematic measurement errors can be greatly minimized. Nonetheless, systematic measurement errors serve as a reminder that it is very difficult, if not impossible, to truly assess a trait in pure isolation from other traits.

3.8 MEASUREMENT ERROR AND RELIABILITY

Perhaps at this point the reader is wondering what measurement error has to do with reliability. The most obvious connection is that measurement error reduces the reliability or repeatability of psychological test results. In fact, we will show here that reliability bears a precise statistical relationship to measurement error. Reliability and measurement error are really just different ways of expressing the same concern: How consistent is a psychological test? The interdependence of these two concepts will become clear if we provide a further sketch of the classical theory of measurement.

A crucial assumption of classical theory is that unsystematic measurement errors act as random influences. This does not mean that the sources of measurement error are completely mysterious and unfathomable in every individual case. We might suspect for one person that her score on digit span reflected a slight negative measurement error caused by the auditory interference of someone coughing in the hallway during the presentation of the fifth item. Likewise, we could conjecture that another person received the benefit of positive measurement error by glimpsing in the mirror behind the examiner to see the correct answer to the ninth item on an information test. Thus, measurement error is not necessarily a mysterious event in every individual case.

However, when we examine the test scores of groups of persons, the causes of measurement error are incredibly complex and varied. In this context, unsystematic measurement errors behave like random variables. The classical theory accepts this essential randomness of measurement error as an axiomatic assumption.

Because they are random events, unsystematic measurement errors are equally likely to be positive or negative and will, therefore, average out to zero across a large group of subjects. Thus, a second assumption is that the mean error of measurement is zero. Classical theory also assumes that measurement errors are not correlated with true scores. This makes intuitive sense: If the error scores were related to another score, it would suggest that they were systematic rather than random, which would violate the essential assumption of classical theory. Finally, it is also assumed that measurement errors are not correlated with errors on other tests.

We can summarize the main features of classical theory as follows (**Gulliksen, 1950 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib672)**, chap. 2):

- Measurement errors are random.
- Mean error of measurement = 0.
- True scores and errors are uncorrelated: $r_{Te} = 0$.
- Errors on different tests are uncorrelated: $r_{12} = 0$.

Starting from these assumptions, it is possible to develop a number of important implications for reliability and measurement. (The points that follow are based on the optimistic assumption that systematic measurement errors are minimal or nonexistent for the instrument in question.) For example, we know that any test administered to a large group of persons will show a variability of obtained scores that can be expressed statistically as a variance, that is, σ^2. The value of classical theory is that it permits us to partition the variance of obtained scores into two separate sources. Specifically, it can be shown that the variance of obtained scores is simply the variance of true scores plus the variance of errors of measurement:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$$

We will refer the interested reader to Gulliksen (**1950 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib672)**, chap. 3) for the computational details.

The preceding formula demonstrates that test scores vary as the result of two factors: variability in true scores, and variability due to measurement error. The obvious implication of this relationship is that errors of measurement contribute to inconsistency of obtained test scores; results will not remain stable if the test is administered again.

3.9 THE RELIABILITY COEFFICIENT

We are finally in a position to delineate the precise relationship between reliability and measurement error. By now the reader should have discerned that reliability expresses the relative influence of true and error scores on obtained test scores. In more precise mathematical terms, the **reliability coefficient (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss275)** (r_{XX}) is the ratio of true score variance to the total variance of test scores. That is:

$$r_{XX} = \frac{\sigma_T^2}{\sigma_X^2}$$

or equivalently:

$$r_{XX} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}$$

Note that the range of potential values for r_{XX} can be derived from analysis of the preceding formula. Consider what happens when the variance due to measurement error (σ_e^2) is very small, close to zero. In that event, the reliability coefficient (r_{XX}) approaches a value of (σ_T^2/σ_T^2) or 1.0. At the opposite extreme, where the variance due to measurement error is very large, the value of the reliability coefficient becomes smaller, approaching a theoretical limit of 0.0. In sum, a completely unreliable test (large measurement error) will yield a reliability coefficient close to 0.0, while a completely reliable test (no measurement error) will produce a reliability coefficient of 1.0. Thus, the possible range of the reliability coefficient is between 0.0 and 1.0. In practice, all tests produce reliability coefficients somewhere in between, but the closer the value of r_{XX} to 1.0, the better.

In a literal sense, r_{XX} indicates the proportion of variance in obtained test scores that is accounted for by the variability in true scores. However, the formula for the reliability coefficient r_{XX} indicates an additional interpretation of it as well. The reader will recall that obtained scores are symbolized by *X*s. In like manner, the subscripts in the symbol for the reliability coefficient signify that r_{XX} is an index of the potential or actual consistency of obtained scores. Thus, tests that capture minimal amounts of measurement error produce consistent and reliable scores; their reliability coefficients are near 1.0. Conversely, tests that reflect large amounts of measurement error produce inconsistent and unreliable scores; their reliability coefficients are closer to 0.0.

Up to this point, the discussion of reliability has been conceptual rather than practical. We have pointed out that reliability refers to consistency of measurement; that reliability is diminished to the extent that errors of measurement dominate the obtained score; and that one statistical index of reliability, the reliability coefficient, can vary between 0.0 and 1.0. But how is a statistical measure of reliability computed? We approach this topic indirectly, first reviewing an essential statistical tool, the correlation coefficient. The reader will discover that the correlation coefficient, a numerical index of the degree of linear relationship between two sets of scores, is an excellent tool for appraising the consistency or repeatability of test scores. We provide a short refresher on the meaning of correlation before proceeding to a summary of methods for estimating reliability.

3.10 THE CORRELATION COEFFICIENT

In its most common application, a **correlation coefficient** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss73) (r) expresses the degree of linear relationship between two sets of scores obtained from the same persons. Correlation coefficients can take on values ranging from −1.00 to +1.00. A correlation coefficient of +1.00 signifies a perfect linear relationship between the two sets of scores. In particular, when two measures have a correlation of +1.00, the rank ordering of subjects is identical for both sets of scores. Furthermore, when arrayed on a scatterplot (**Figure 3.8a** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec10#ch03fig8)), the individual data points (each representing a pair of scores from a single subject) conform to a perfectly straight line with an upward slope. A correlation coefficient of −1.00 signifies an equally strong relationship but with inverse correspondence: the highest score on one variable corresponding to the lowest score on the other, and vice versa. In this case, the individual data points conform to a perfectly straight line with a downward slope (**Figure 3.8b** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec10#ch03fig8)). Correlations of +1.00 or −1.00 are extremely rare in psychological research and usually signify a trivial finding. For example, if on two occasions in quick succession we counted the number of letters in the last name of 100 students, these two sets of "scores" would show a correlation of +1.00.

Negative correlations usually result from the manner in which one of the two variables was scored. For example, scores on the Category Test (**Reitan & Wolfson, 1993** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1350)) are reported as errors, whereas results on the Raven Progressive Matrices (**Raven, Court, & Raven, 1983** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1338) , **1986** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1339)) are reported as number of items correct. Persons who obtain a high score on the Category Test (many errors) will most likely obtain a low score on the Progressive Matrices test (few correct). Thus, we would expect a substantial negative correlation for scores on these two tests.

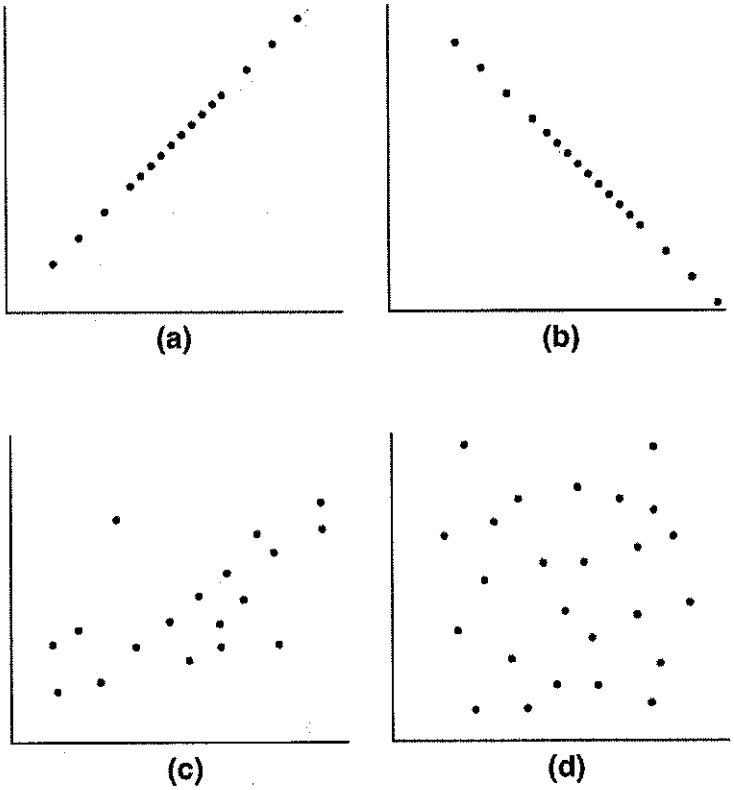


FIGURE 3.8 Scatterplots Depicting Different Degrees of Correlation

Consider the scatterplot in **Figure 3.8c** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec10#ch03fig8) , which might depict the hypothetical heights and weights of a group of persons. As the reader can see, height and weight are strongly but not perfectly related to one another. Tall persons tend to weigh more, short persons less, but there are some exceptions. If we were to compute the correlation coefficient between height and weight—a simple statistical task outlined in the following—we would obtain a value of about +.80, indicating a strong, positive relationship between these measures.

When two variables have no relationship, the scatterplot takes on an undefined bloblike shape and the correlation coefficient is close to 0.00 (**Figure 3.8d** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec10#ch03fig8)). For example, in a sample of adults, the correlation between reaction time and weight would most likely be very close to zero.

Finally, it is important to understand that the correlation coefficient is independent of the mean. For example, a correlation of +1.00 can be found between two administrations of the same test even when there are significant mean differences between pretest and posttest. In sum, perfect correlation does not imply identical pre- and posttest scores for each examinee. However, perfect correlation does imply perfectly ordered ranking from pretest to posttest, as discussed previously.

3.11 THE CORRELATION COEFFICIENT AS A RELIABILITY COEFFICIENT

One use of the correlation coefficient is to gauge the consistency of psychological test scores. If test results are highly consistent, then the scores of persons taking the test on two occasions will be strongly correlated, perhaps even approaching the theoretical upper limit of +1.00. In this context, the correlation coefficient is also a reliability coefficient. Even though the computation of the Pearson *r* makes no reference to the theory of true and error scores, the correlation coefficient does, nonetheless, reflect the proportion of variance in obtained test scores accounted for by the variability in true scores. Thus, in some contexts a correlation coefficient is a reliability coefficient.

This discussion introduces one method for estimating the reliability of a test: Administer the instrument twice to the same group of persons and compute the correlation between the two sets of scores. The test–retest approach is very common in the evaluation of reliability, but several other strategies exist as well. As we review the following methods for estimating reliability, the reader may be temporarily bewildered by the apparent diversity of approaches. In fact, the different methods fall into two broad groups, namely, temporal stability approaches, which directly measure the consistency of test scores, and internal consistency approaches, which rely upon a single test administration to gauge reliability. Keep in mind that one common theme binds all the eclectic methods together: Reliability is always an attempt to gauge the likely accuracy or repeatability of test scores.

3.12 RELIABILITY AS TEMPORAL STABILITY

Test–Retest Reliability

The most straightforward method for determining the reliability of test scores is to administer the identical test twice to the same group of heterogeneous and representative subjects. If the test is perfectly reliable, each person's second score will be completely predictable from his or her first score. On many kinds of tests, particularly ability and achievement tests, we might expect subjects generally to score somewhat higher the second time because of practice, maturation, schooling, or other intervening effects that take place between pretest and posttest. However, so long as the second score is strongly correlated with the first score, the existence of practice, maturation, or treatment effects does not cast doubt on the **test–retest reliability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss331) of a psychological test.

An example of a reliability coefficient computed as a test-retest correlation coefficient is depicted in **Figure 3.9** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec12#ch03fig9) . In this case, 60 subjects were administered the Finger Tapping Test (FTT) on two occasions separated by a week (**Morrison, Gregory, & Paul, 1979** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1178)). The FTT, one component of the Halstead-Reitan neuropsychological test battery (**Reitan & Wolfson, 1993** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1350)), is a relatively pure measure of motor speed. Using a standardized mechanical counting apparatus, the subject is instructed to tap with the index finger as fast as possible for 10 seconds. This procedure is continued until five trials in a row reveal consistent results. The procedure is repeated for the nondominant hand. The score for each hand is the average of the five consecutive trials.

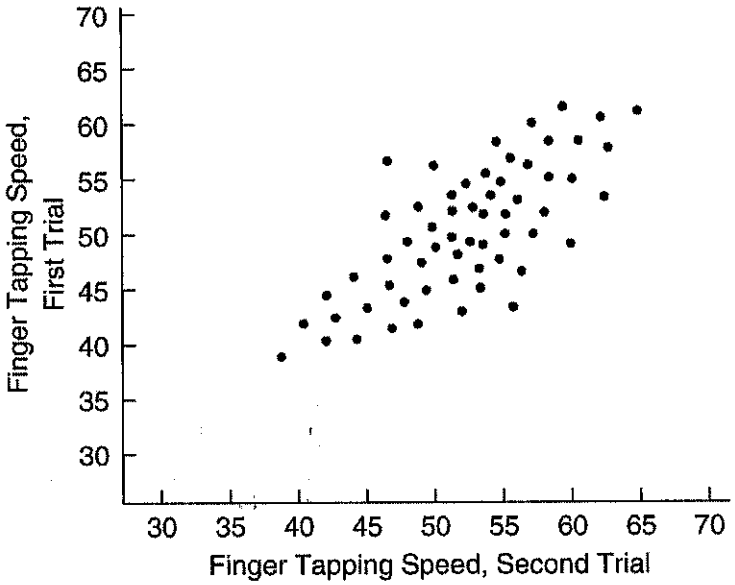


FIGURE 3.9 Scatterplots Revealing a Reliability Coefficient of .80

Source: Based on data from Morrison, M. W., Gregory, R. J., & Paul, J. J. (1979). Reliability of the Finger Tapping Test and a note on sex differences. *Perceptual and Motor Skills, 48,* 139–142.

The correlation between scores from repeated administrations of this test works out to be about .80. This is at the low end of acceptability for reliability coefficients, which usually fall in the .80*s* or .90*s*. We discuss standards of reliability in more detail subsequently.

Alternate-Forms Reliability

In some cases test developers produce two forms of the same test. These alternate forms are independently constructed to meet the same specifications, often on an item-by-item basis. Thus, alternate forms of a test incorporate similar content and cover the same range and level of difficulty in items. Alternate forms of a test possess similar statistical and normative properties. For example, when administered in counterbalanced fashion to the same group of subjects, the means and standard deviations of alternate forms are typically quite comparable.

Estimates of **alternate-forms reliability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss08) are derived by administering both forms to the same group and correlating the two sets of scores. This approach has much in common with test–retest methods—both strategies involve two test administrations to the same subjects with an intervening time interval. For both approaches, we would expect that intervening changes in motivation and individual differences in amount of improvement would produce fluctuations in test scores and thereby reduce reliability estimates somewhat. Thus, test–retest and alternate-forms reliability estimates share considerable conceptual similarity.

However, there is one fundamental difference between these two approaches. The alternate-forms methodology introduces item-sampling differences as an additional source of error variance. That is, some test takers may do better or worse on one form of a test because of the particular items sampled. Even though the two forms may be equally difficult on average, some subjects may find one form quite a bit harder (or easier) than the other because supposedly parallel items are not equally familiar to every person. Notice that item-sampling differences are not a source of error variance in the test-retest approach because identical items are used in both administrations.

Alternate forms of a test are also quite expensive—nearly doubling the cost of publishing a test and putting it on the market. Because of the increased cost and also the psychometric difficulties of producing truly parallel forms, fewer and fewer tests are being released in this format.

3.13 RELIABILITY AS INTERNAL CONSISTENCY

We turn now to some intriguing ways of estimating the reliability of an individual test without developing alternate forms and without administering the test twice to the same examinees (**Feldt & Brennan, 1989** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib495)). The first approach correlates the results from one-half of the test with the other half and is appropriately termed split-half reliability. The second approach examines the internal consistency of individual test items. In this method, the psychometrician seeks to determine whether the test items tend to show a consistent interrelatedness. Finally, insofar as some tests are less than perfectly reliable because of differences among scorers, we also take up the related topic of interscorer reliability.

Split-Half Reliability

We obtain an estimate of **split-half reliability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss302) by correlating the pairs of scores obtained from equivalent halves of a test administered only once to a representative sample of examinees. The logic of split-half reliability is straightforward: If scores on two half tests from a *single* test administration show a strong correlation, then scores on two whole tests from two *separate* test administrations (the traditional approach to evaluating reliability) also should reveal a strong correlation.

Psychometricians typically view the split-half method as supplementary to the gold standard approach, which is the test–retest method. For example, in the standardization of the WAIS-IV, the reliability of most scales was established by the test–retest approach *and* the split-half approach. These two estimates of reliability are generally similar, although split-half approaches often yield higher estimates of reliability.

One justification for the split-half approach is that logistical problems or excessive cost may render it impractical to obtain a second set of test scores from the same examinees. In this case, a split-half estimate of reliability is the only thing available, and it is certainly better than no estimate at all. Another justification for the split-half approach is that the test–retest method is potentially misleading in certain cases. For example, some ability tests are prone to large but inconsistent practice effects—such as when examinees learn concepts from feedback given as part of the standardized testing procedure. When practice effects are large and variable, the rank order of scores from a second administration will at best sustain only a modest association to the rank order of scores from the first administration. For these kinds of instruments, test–retest reliability coefficients could be misleadingly low. Finally, test–retest approaches also will yield misleadingly low estimates of reliability if the trait being measured is known to fluctuate rapidly (e.g., certain measures of mood).

The major challenge with split-half reliability is dividing the test into two nearly equivalent halves. For most tests—especially those with the items ranked according to difficulty level—the first half is easier than the second half. We would not expect examinees to obtain equivalent scores on these two portions, so this approach to splitting a test rarely is used. The most common method for obtaining split halves is to compare scores on the odd items versus the even items of the test. This procedure works particularly well when the items are arranged in approximate order of difficulty.

In addition to calculating a Pearson *r* between scores on the two equivalent halves of the test, the computation of a coefficient of split-half reliability entails an additional step: adjusting the half-test reliability using the Spearman-Brown formula.

The Spearman-Brown Formula

Notice that the split-half method gives us an estimate of reliability for an instrument half as long as the full test. Although there are some exceptions, a shorter test generally is less reliable than a longer test. This is especially true if, in comparison to the shorter test, the longer test embodies equivalent content and similar item difficulty. Thus, the Pearson *r* between two halves of a test will usually underestimate the reliability of the full instrument. We need a method for deriving the reliability of the whole test based on the half-test correlation coefficient.

The **Spearman-Brown formula** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss298) provides the appropriate adjustment:

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

In this formula, r_{SB} is the estimated reliability of the full test computed by the Spearman-Brown method, while r_{hh} is the half-test reliability. **Table 3.8** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec13#ch03tab8) shows conceivable half-test correlations alongside the corresponding Spearman-Brown reliability coefficients for the whole test. For example, using the Spearman-Brown formula, we could determine that a half-test reliability of .70 is equivalent to an estimated full-test reliability of .82.

TABLE 3.8 Comparison of Split-Half Reliabilities and Corresponding Spearman-Brown Reliabilities

Split-Half Reliability	*Spearman-Brown Reliability*
.5	.67
.6	.75
.7	.82
.8	.89
.9	.95

Critique of the Split-Half Approach

Although the split-half approach is widely used, nonetheless it has been criticized for its lack of precision:

> Instead of giving a single coefficient for the test, the procedure gives different coefficients depending on which items are grouped when the test is split into two parts. If one split may give a higher coefficient than another, one can have little faith in whatever result is obtained from a single split. (**Cronbach, 1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib373))

Why rely on a single split? Why not take a more typical value such as the mean of the split-half coefficients resulting from all possible splittings of a test? Cronbach (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib373)) advocated just such an approach when proposing a general formula for estimating the reliability of a psychological test.

Coefficient Alpha

As proposed by Cronbach (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib373)) and subsequently elaborated by others (**Novick & Lewis, 1967; Kaiser & Michael, 1975** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib851)), **coefficient alpha** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss52) may be thought of as the mean of all possible split-half coefficients, corrected by the Spearman-Brown formula. The formula for coefficient alpha is

$$r_\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\Sigma\sigma_j^2}{\sigma^2}\right)$$

where r_α is the coefficient alpha, N is the number of items, σ_j^2 is the variance of one item, $\Sigma\sigma_j^2$ is the sum of variances of all items, and σ^2 is the variance of the total test scores. As with all reliability estimates, coefficient alpha can vary between 0.00 and 1.00.

Coefficient alpha is an index of the internal consistency of the items, that is, their tendency to correlate positively with one another. Insofar as a test or scale with high internal consistency will also tend to show stability of scores in a test–retest approach, coefficient alpha is therefore a useful estimate of reliability.

Traditionally, coefficient alpha has been thought of as an index of unidimensionality, that is, the degree to which a test or scale measures a single factor. Recent analyses by Schmitt (**1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1469)) serve to dispel this misconception. Certainly coefficient alpha is an index of the interrelatedness of the individual items, but this is not synonymous with the unidimensionality of what the test or scale measures. In fact, it is possible for a scale to measure two or more distinct factors and yet still possess a very strong coefficient alpha. Schmitt (**1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1469)) gives the example of a six-item test in which the first three items correlate .8 one with another, the last three items also correlate .8 one with another, whereas correlations across the two three-item sets are only .3 (**Table 3.9** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec13#ch03tab9)). Even though this is irrefutably a strong two-factor test, the value for coefficient alpha works out to be 86! For this kind of test, coefficient alpha probably will overestimate test–retest reliability. This is why psychometricians look to test–retest approaches as essential to the evaluation of reliability. Certainly the split-half approach in general and coefficient alpha in particular are valuable approaches to reliability, but they cannot replace the common sense of the test–retest approach: When the same test is administered twice to a representative sample of examinees, do they obtain the same relative placement of scores?

TABLE 3.9 A Six-Item Test with Two Factors and Strong Coefficient Alpha

Variable	*1*	*2*	*3*	*4*	*5*	*6*
1	—					
2	.8	—				
3	.8	.8	—			
4	.3	.3	.3	—		
5	.3	.3	.3	.8	—	
6	.3	.3	.3	.8	.8	—

Note: Coefficient alpha = .86.

Source: Reprinted with permission from Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350–353.

The Kuder-Richardson Estimate of Reliability

Cronbach (**1951** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib373)) has shown that coefficient alpha is the general application of a more specific formula developed earlier by Kuder and Richardson (**1937** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib928)). Their formula is generally referred to as **Kuder-Richardson formula 20** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss180) or, simply, KR-20, in reference to the fact that it was the twentieth in a lengthy series of derivations. The KR-20 formula is relevant to the special case in which each test item is scored 0 or 1 (e.g., wrong or right). The formula is

$$\text{KR-20} = \left(\frac{N}{N-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

where

- N = the number of items on the test,
- σ^2 = the variance of scores on the total test,
- p = the proportion of examinees getting each item correct,
- q = the proportion of examinees getting each item wrong.

Coefficient alpha extends the Kuder-Richardson method to types of tests with items that are not scored as 0 or 1. For example, coefficient alpha could be used with an attitude scale in which examinees indicate on each item whether they strongly agree, agree, disagree, or strongly disagree **interscorer reliability** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss164).

Interscorer Reliability

Some tests leave a great deal of judgment to the examiner in the assignment of scores. Certainly, projective tests fall into this category, as do tests of moral development and creativity. Insofar as the scorer can be a major factor in the reliability of these instruments, a report of interscorer reliability is imperative. Computing is a very straightforward procedure. A sample of tests is independently scored by two or more examiners and scores for pairs of examiners are then correlated. Test manuals typically report the training and experience required of examiners and then list representative interscorer correlation coefficients.

Interscorer reliability supplements other reliability estimates but does not replace them. It would still be appropriate to assess the test–retest or other type of reliability in a subjectively scored test. We provide a quick summary of methods for estimating reliability in **Table 3.10** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec13#ch03tab10).

Which Type of Reliability Is Appropriate?

As noted, even when a test has only a single form, there are still numerous methods available for assessing reliability: test–retest, split-half, coefficient alpha, and interscorer methods. For tests that possess two forms, we can add a fifth method: alternate-forms reliability. Which method is best? When should we use one method but not another? To answer these questions, we need to know the nature and purpose of the individual test in question.

For tests designed to be administered to individuals more than once, it would be reasonable to expect that the test demonstrate reliability across time—in this case, test–retest reliability is appropriate. For tests that purport to possess factorial purity, coefficient alpha would be essential. In contrast, factorially complex tests such as measures of general intelligence would not fare well by measures of internal consistency. Thus, coefficient alpha is not an appropriate index of reliability for all tests but applies only to measures that are designed to assess a single factor. Split-half methods work well for instruments that have items carefully ordered according to difficulty level. Of course, interscorer reliability is appropriate for any test that involves subjectivity of scoring.

It is common for test manuals to report multiple sources of information about reliability. For example, the WAIS-IV Manual **(Wechsler, 2008 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1737))** reports split-half reliabilities for most subtests and also provides test–retest coefficients for all subtests and IQ scores. The manual also cites information akin to alternate-forms reliability—it reports the correlations between the WAIS-IV and its predecessor, the WAIS-III.

In order to analyze the error variance into its component parts, a number of reliability coefficients will need to be computed. Although it is difficult to arrive at precise data in the real world, on a theoretical basis we can partition the variability of scores into true and error components as depicted in **Figure 3.10 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec13#ch03fig10)** .

TABLE 3.10 Brief Synopsis of Methods for Estimating Reliability

Method	*No. Forms*	*No. Sessions*	*Sources of Error Variance*
Test–Retest	1	2	Changes over time
Alternate-Forms (immediate)	2	1	Item sampling
Alternate-Forms (delayed)	2	2	Item sampling Changes over time
Split-Half	1	1	Item sampling Nature of split
Coefficient Alpha	1	1	Item sampling Test heterogeneity
Interscorer	1	1	Scorer differences

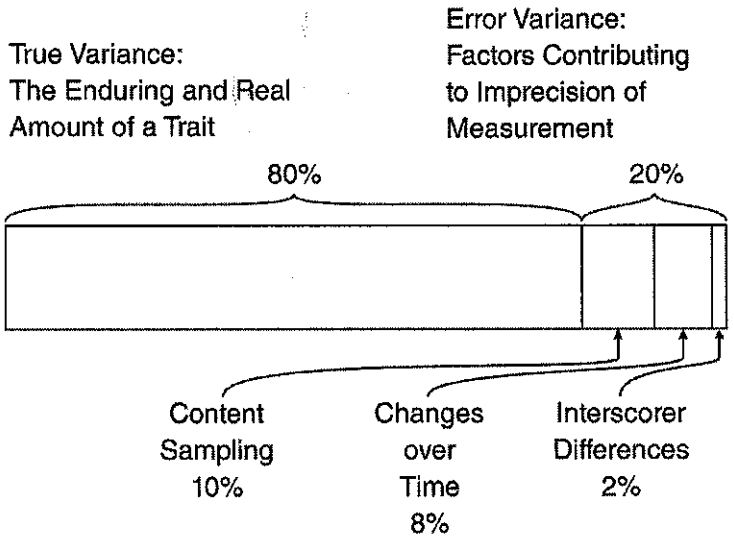


FIGURE 3.10 Sources of Variance in a Hypothetical Test

Note: The results are similar to what might be found if alternative forms of an individual intelligence test were administered to the same person by different examiners.

3.14 ITEM RESPONSE THEORY

The classical test theory summarized previously dominated test development for most of the twentieth century. However, beginning slowly in the 1960s and continuing to the present time, psychometricians have favored an alternative model of test theory known as **item response theory** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss176) (IRT) or **latent trait theory** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss182) (**Embretson, 1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib458) ; **Lord & Novick, 1968** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1006) ; **Rasch, 1960** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1333)). IRT is more than a theory; it is also a collection of mathematical models and statistical tools with widespread uses. The applications of IRT include analyzing items and scales, developing homogeneous psychological measures, measuring individuals on psychological constructs (e.g., depression, intelligence, leadership), and administering psychological tests by computer. The foundational elements of IRT include item response functions (IRFs), information functions, and the assumption of invariance (**Reise, Ainsworth, & Haviland, 2005** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1348)).

Item Response Functions

An **item response function** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss175) (IRF), also known as an item characteristic curve (ICC), is a mathematical equation that describes the relation between the amount of a latent trait an individual possesses and the probability that he or she will give a designated response to a test item designed to measure that construct. In the case of ability measures, the designated response is the correct answer, whereas in other situations (e.g., the measurement of personality constructs such as leadership), the designated response would be the one indicating the presence of the trait being assessed. For the sake of simplicity, we will refer to the designated response as the "correct" response in the discussion that follows.

Each respondent is assumed to have a certain amount of the latent trait being measured, whether this is verbal proficiency, spatial memory, or leadership ability. In turn, the latent trait is assumed to influence directly the examinee's responses to the items on the test, which has been carefully designed to measure the trait in question. The mathematical models and statistical tools of IRT are designed to establish the IRF for each item on the test. Collectively, the IRFs can be used for many purposes, including the refinement of the instrument, the calculation of reliability, and the estimation of examinee trait levels. For example, test developers commonly use IRFs to eliminate items that don't function optimally in a psychometric sense.

Each test item has its own IRF. The IRFs for four dichotomously scored items are plotted in **Figure 3.11** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig11) . The trait level is depicted on the abscissa, with standard scores ranging from −3 to +3. An average amount of the trait in question would be indicated by a score of 0. Actually, for mathematical reasons, the scores in an IRF can range hypothetically from −∞ to +∞, but in actual practice, scores rarely escape the bounds of −3 to +3. The ordinate depicts the probability of a correct response on a scale from 0 to 1.

Upon careful reflection, the IRF provides a wealth of information about each item. For example, it can be used to determine the difficulty level of test items. In the IRT approach, difficulty level is gauged differently than in classical test theory. According to classical test theory, the difficulty level of an item is equivalent to the proportion of examinees in a standardization sample who pass the item. In contrast, according to IRT, difficulty is indexed by how much of the trait is needed to answer the item correctly. For the items shown in **Figure 3.11** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig11) , item A has the lowest difficulty level—it is passed by almost everyone, even examinees possessing only a small amount of the trait in question. In contrast, item D has the highest difficulty level—only those with high amounts of the trait typically answer correctly. Although not immediately obvious, items B and C are equal in difficulty level—for example, individuals with an average trait level (a score of 0) have a 50 percent chance of answering these items correctly.

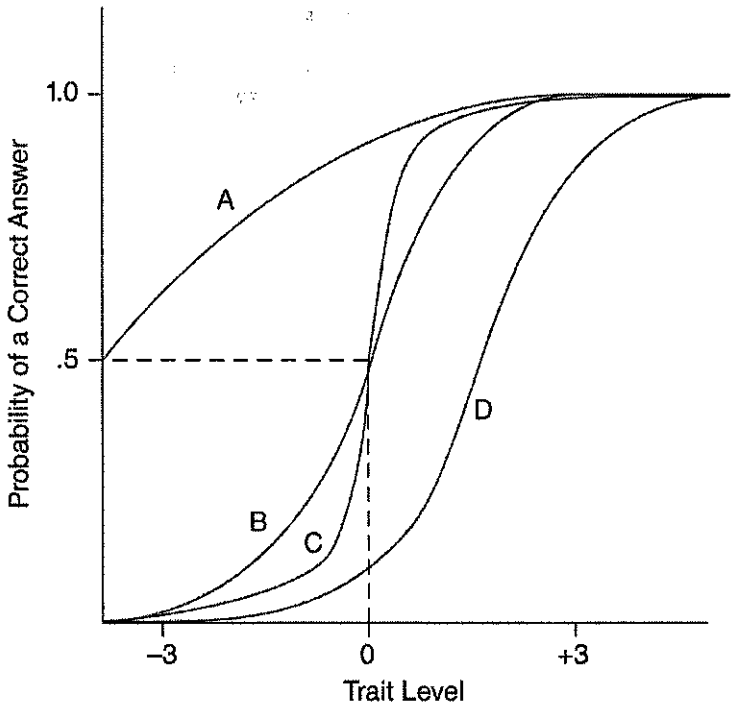


FIGURE 3.11 Item Response Functions for Four Test Items

Another quality evident in the IRF is the item discrimination parameter, which is a gauge of how well the item differentiates among individuals at a specific level of the trait in question. Consider items B and C in **Figure 3.11** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig11). Although they are equally difficult overall—both answered correctly by 50 percent of the examinees—item C with its steeper curve possesses better discrimination, meaning that it is better able to differentiate among individuals at this level of the trait.

The appealing advantage of the IRT approach to measurement is that the probability of a respondent answering a particular question correctly can be expressed as a precise mathematical equation. Although it is beyond the scope of our presentation to go into the derivation, seeing an IRT equation might help the reader appreciate the sophistication of this approach. We denote the item difficulty as b and the amount of the trait that an examinee possesses as θ. Then the relevant equation looks like this:

$$p(\theta) = 1/(1 + e^{-(\theta - b)})$$

where $p(\theta)$ is the probability of a respondent with trait level θ correctly responding to an item of difficulty b. When the parameters are filled in and this equation is plotted, the outcome is an IRF for each test item, similar to those shown in **Figure 3.11** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig11). The symbol e in the equation refers to the base for natural logarithms, which has a constant value of 2.71828. The parameter θ refers to the examinee trait level measured on a standard scale, which typically varies from −3 to +3. This particular formula was developed by the Danish mathematician Georg Rasch (**1960** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1333)); hence, in his honor this IRT application is also known as a **Rasch Model** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss267). This is a simple and elegant application of IRT, also known as the one parameter model. The single parameter referred to is b, the item difficulty level. More complex models have also been developed. These include the two-parameter model that adds the item discrimination index to the equation, and the three-parameter model that factors in a guessing parameter as well (**Baker, 2001** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib78)). The discussion here is based on the one-parameter model.

Information Functions

In general terms, information is that which reduces uncertainty. In psychological measurement, information represents the capacity of a test item to differentiate among people (**Reise, Ainsworth, & Haviland, 2005** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1348)). On most scales, certain items are intended to differentiate among individuals low on the trait being measured, whereas other items are designed for discrimination at higher trait levels. Consider items A and D from **Figure 3.11** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig11). Item A is useful only for testing individuals low on the relevant trait—at higher levels, everyone answers correctly, and no information is gained. It would be pointless to administer this item to individuals at the higher end of the trait spectrum because it is certain they will answer correctly. Conversely, item D is useful only for individuals with high trait levels—at lower trait levels, it is certain that everyone fails the item and, likewise, no information is gained.

Another way of stating this is to say that a test item typically provides a different level of information at each level of the trait in question. For example, item A provides a lot of information at low trait levels but none at high levels, whereas item D shows the reverse pattern—no information at low trait levels but more information at high levels. Using a simple mathematical conversion, an **item information function** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss173) can be derived from the IRF for each item. This function portrays graphically the relationship between the trait level of examinees and the information provided by the test item. The information functions for items A and D are displayed in **Figure 3.12** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec14#ch03fig12).

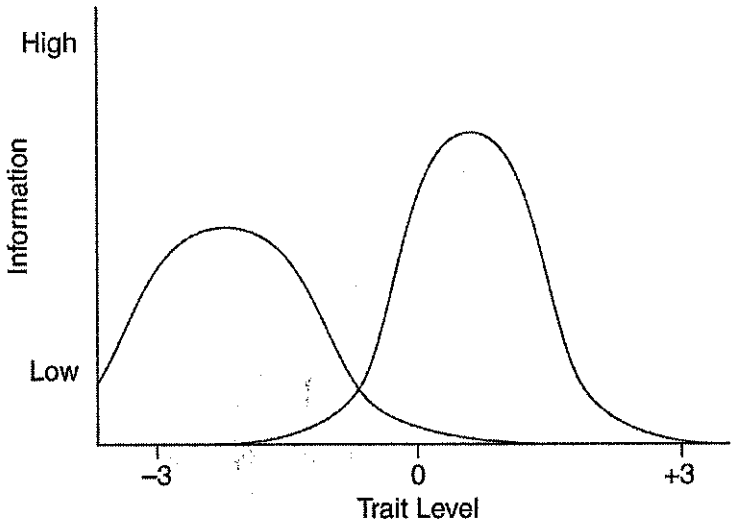


FIGURE 3.12 Item Information Functions for Two Test Items

The beauty of IRT is that the item information functions from different scale items can be *added together* to derive the scale information function:

> Because information is directly related to measurement precision (more information equals more precise measurement), the scale information function estimates how well a measure functions as a whole in different trait ranges. The fact that item information functions can be added together is the foundation for scale construction with IRT. (**Reise, Ainsworth, & Haviland, 2005** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1348), p. 96)

The scale information function is analogous to test reliability as elucidated in classical test theory with two important differences. First, in IRT the precision of measurement can vary, depending on where an individual falls in the trait range, whereas in classical test theory a single reliability (precision of measurement) is typically calculated for the entire test. Second, in IRT a different collection of test items might be used for each examinee to obtain a predetermined precision of measurement, whereas in classical test theory a single set of items is typically administered to all examinees.

Invariance in IRT

Invariance is a challenging concept to understand because it is contrary to the traditional lore of testing, which posits that test scores are meaningful only in a relative sense—in relation to fixed scales administered to large standardization samples. Certainly, it is true within IRT that huge databases are needed to make sense of individual test results. Yet, within IRT the manner in which we estimate the trait level (i.e., acquire a score) is fundamentally different from traditional approaches such as classical test theory.

Within the IRT framework, invariance refers to two separate but related ideas **(Reise, Ainsworth, & Haviland, 2005 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1348)**). First, invariance means that an examinee's position on a latent-trait continuum (his or her score) can be estimated from the responses to any set of test items with known IRFs. In other words, as long as the IRFs for a particular set of test items have been previously calculated, a trait level can be estimated for an examinee who has answered those items. In fact, the particular items used might differ from one examinee to another, and the number of items administered might even differ. But as long as the IRFs of the particular items are known, the methods of IRT provide an estimate of the trait level (i.e., a test score). Preferably, of course, items with appropriate difficulty levels corresponding to the trait level of the examinee will be administered. Typically, this is accomplished by using computer programs that flexibly select test items based on the prior responses of the examinee.

The second meaning of invariance is that the IRFs do not depend on the characteristics of a particular population. In other words, the IRF for each item is presumed to exist in some abstract, independent, and enduring manner, waiting to be discovered by the psychometrician. The results for different samples might help fine-tune different parts of the IRF, but the outcome always should fall on the same curve. This means, as well, that the scale of the trait exists independently of any set of items and independently of any particular population. Reise, Ainsworth, and Haviland **(2005 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1348)**) describe the advantages of item-parameter invariance as follows:

> For example, in large-scale educational assessment, item-parameter invariance facilitates the linking of scales from different measures (i.e., placing scores on a single, common scale), across students in different grade levels (e.g., third through sixth grade in the same school) and within a grade level (e.g., fourth graders in different schools). Similarly, using IRT methods to compare individuals who have responded to different measures is relevant to cross-cultural and developmental researchers. . . . (p. 98)

Although IRT analyses typically require large samples—several hundred or thousands of respondents—the necessary software is straightforward and commonly available. Given its advantages, IRT approaches to test development likely will become increasingly prominent in the years ahead.

3.15 THE NEW RULES OF MEASUREMENT

When fully explicated, IRT leads to what Embretson (**1996** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib458)) has called "the new rules of measurement." By this she means that several conclusions from classical testing theory do not hold true within the framework of IRT. For example, within classical testing theory, the standard error of measurement is assumed to be a constant that applies to all examinee scores regardless of the ability level of a particular respondent. However, within IRT the standard error of measurement becomes substantially larger at both extremes of ability. In other words, the IRT model concludes that test scores are more reliable for individuals of average ability and increasingly less reliable for those with very high or very low ability.

Another difference pertains to the relationship between test length and reliability. In classical test theory, it is almost an axiom that longer tests are more reliable than shorter tests. For example, this follows from the Spearman-Brown formula discussed earlier in the chapter. However, when IRT models are used, shorter tests can be more reliable than longer tests. This is especially true when there is a good match between the difficulty level of the specific items administered and the proficiency level of the examinee. A good fit between these two parameters allows for a precise (reliable) estimate of ability using a relatively smaller number of test items.

In general, tests developed within an IRT model are better suited to computerized adaptive testing, in which a computer program is used not only to administer test items but also to select them in a flexible manner based on each examinee's ongoing responses to prior items. Computerized adaptive testing is discussed in more detail in **Topic 12B** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch12lev1sec5#ch12box3) , Computerized Assessment and the Future of Testing.

3.16 SPECIAL CIRCUMSTANCES IN THE ESTIMATION OF RELIABILITY

Traditional approaches to estimating reliability may be misleading or inappropriate for some applications. Some of the more problematic situations involve unstable characteristics, speed tests, restriction of range, and criterion-referenced tests.

Unstable Characteristics

Some characteristics are presumed to be ever changing in reaction to situational or physiological variables. Emotional reactivity as measured by electrodermal or galvanic skin response is a good example. Such a measure fluctuates quickly in reaction to loud noises, underlying thought processes, and stressful environmental events. Even just talking to another person can arouse a strong electrodermal response. Because the true amount of emotional reactivity changes so quickly, test and retest must be nearly instantaneous in order to provide an accurate index of reliability for unstable characteristics such as an electrodermal measure of emotional reactivity.

Speed and Power Tests

A speed test typically contains items of uniform and generally simple levels of difficulty. If time permitted, most subjects should be able to complete most or all of the items on such a test. However, as the name suggests, a speeded test has a restrictive time limit that guarantees few subjects complete the entire test. Since the items attempted tend to be correct, an examinee's score on a speeded test largely reflects speed of performance.

Speed tests are often contrasted with power tests. A power test allows enough time for test takers to attempt all items but is constructed so that no test taker is able to obtain a perfect score. Most tests contain a mixture of speed and power components.

The most important point to stress about the reliability of speed tests is that the traditional split-half approach (comparing odd and even items) will yield a spuriously high reliability coefficient. Consider one test taker who completes 60 of 90 items on a speed test. Most likely, the odd-even approach would show 30 odd items correct and 30 even items correct. With similar data from other subjects, the correlation between scores on odd and even items necessarily would approach +1.00. The reliability of a speed test should be based on the test–retest method or split-half reliability from two, separately timed half tests. In the latter instance, the Spearman-Brown correction is needed.

Restriction of Range

Test–retest reliability will be spuriously low if it is based on a sample of homogeneous subjects for whom there is a **restriction of range** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss277)** on the characteristic being measured. For example, it would be inappropriate to estimate the reliability of an intelligence test by administering it twice to a sample of college students. This point is illustrated by the hypothetical but realistic scatterplot shown in **Figure 3.13** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec16#ch03fig13)** , where the reader can see a strong test–retest correlation for the entire range of diverse subjects, but a weak correlation for brighter subjects viewed in isolation.

Reliability of Criterion-Referenced Tests

The reader will recall from the first topic of this chapter that criterion-referenced tests evaluate performance in terms of mastery rather than assessing a continuum of achievement. Test items are designed to identify specific skills that need remediation; therefore, items tend to be of the "pass/fail" variety.

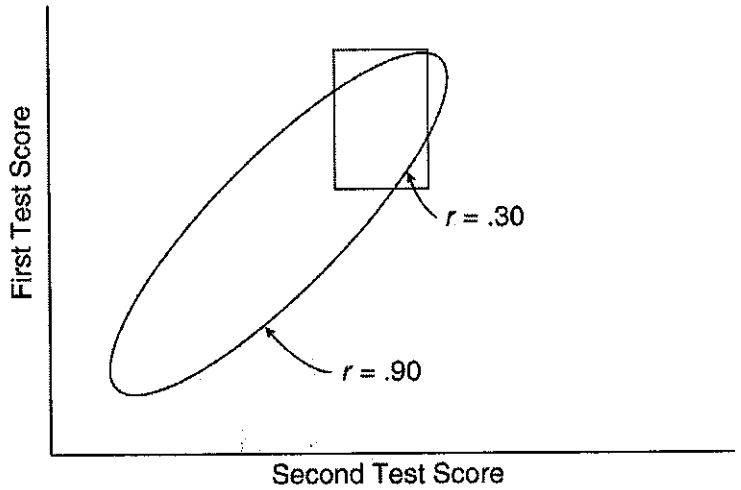


FIGURE 3.13 Sampling a Restricted Range of Subjects Causes Test-Retest Reliability to Be Spuriously Low

The structure of criterion-referenced tests is such that the variability of scores among examinees is typically quite minimal. In fact, if test results are used for training purposes and everyone continues training until all test skills are mastered, variability in test scores becomes nonexistent. Under these conditions, traditional approaches to the assessment of reliability are simply inappropriate.

With many criterion-referenced tests, results must be almost perfectly accurate to be useful. For example, any classification error is serious if the purpose of a test is to determine a subject's ability to drive a manual transmission, or stick shift, automobile. The key issue here is not whether test and retest scores are close to one another, but whether the classification ("can do/can't do") is the same in both instances. What we really want to know is the percentage of persons for whom the same decision is reached on both occasions—the closer to 100 percent, the better. This is but one illustration of the need for specialized techniques in the evaluation of nonnormative tests. Berk **(1984** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib148)**) and Feldt and Brennan **(1989** **(http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib495)**) discuss approaches to the reliability of criterion-referenced tests.

3.17 THE INTERPRETATION OF RELIABILITY COEFFICIENTS

The reader should now be well versed in the different approaches to reliability and should possess at least a conceptual idea of how reliability coefficients are computed. In addition, we have discussed the distinctive testing conditions that dictate the use of one kind of reliability method as opposed to others. No doubt, the reader has noticed that we have yet to discuss one crucial question: What is an acceptable level of reliability?

Many authors suggest that reliability should be at least .90 if not .95 for decisions about individuals (**e.g., Nunnally & Bernstein, 1994** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1244)). However, there is really no hard and fast answer to this question. We offer the loose guidelines suggested by Guilford and Fruchter (**1978** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib668)):

> There has been some consensus that to be a very accurate measure of individual differences in some characteristic, the reliability should be above .90. The truth is, however, that many standard tests with reliabilities as low as .70 prove to be very useful. And tests with reliabilities lower than that can be useful in research.

On a more practical level, acceptable standards of reliability hinge on the amount of measurement error the user can tolerate in the proposed application of a test. Fortunately, reliability and measurement error are mutually interdependent concepts. Thus, if the test user can specify an acceptable level of measurement error, then it is also possible to determine the minimum standards of reliability required for that specific application of a test. We pursue this topic further by introducing a new concept: standard error of measurement.

3.18 RELIABILITY AND THE STANDARD ERROR OF MEASUREMENT

To introduce the concept of standard error of measurement we begin with a thought experiment. Suppose we could administer thousands of equivalent IQ tests to one individual. Suppose further that each test session was a fresh and new experience for our cooperative subject; in this hypothetical experiment, practice and boredom would have no effect on later test scores. Nonetheless, because of the kinds of random errors discussed in this chapter, the scores of our hapless subject would not be identical across test sessions. Our examinee might score a little worse on one test because he stayed up late the night before; the score on another test might be better because the items were idiosyncratically easy for him. Even though such error factors are random and unpredictable, it follows from the classical theory of measurement that the obtained scores would fall into a normal distribution with a precise mean and standard deviation. Let us say that the mean of the hypothetical IQ scores for our subject worked out to be 110, with a standard deviation of 2.5.

In fact, the mean of this distribution of hypothetical scores would be the estimated true score for our examinee. Our best estimate, then, is that our subject has a true IQ of 110. Furthermore, the standard deviation of the distribution of obtained scores would be the **standard error of measurement** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss304) (SEM). Note that while the true score on a test likely differs from one person to the next, the SEM is regarded as constant, an inherent property of the test. If we repeated this hypothetical experiment with another subject, the estimated true score would probably differ, but the SEM should work out to be a similar value.[2] (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec18#ch03fn02)

As its name suggests, the SEM is an index of measurement error that pertains to the test in question. In the hypothetical case in which *SEM* = 0, there would be no measurement error at all. A subject's obtained score would then also be his or her true score. However, this outcome is simply impossible in real-world testing. Every test exhibits some degree of measurement error. The larger the SEM, the greater the typical measurement error. However, the accuracy or inaccuracy of any individual score is always a probabilistic matter and never a known quantity.

As noted, the SEM can be thought of as the standard deviation of an examinee's hypothetical obtained scores on a large number of equivalent tests, under the assumption that practice and boredom effects are ruled out. Like any standard deviation of a normal distribution, the SEM has well-known statistical uses. For example, 68 percent of the obtained scores will fall within one SEM of the mean, just as 68 percent of the cases in a normal curve fall within one SD of the mean.

The reader will recall from earlier in this chapter that about 95 percent of the cases in a normal distribution fall within two SDs of the mean. For this reason, if our examinee were to take one more IQ test, we could predict with 95 percent odds that the obtained score would be within two SEMs of the estimated true IQ of 110. Knowing that the SEM is 2.5, we would therefore predict that the obtained IQ score would be 110 ± 5; that is, the true score would very likely (95 percent odds) fall between 105 and 115.

Unfortunately, in the real world we do not have access to true scores and we most certainly cannot obtain multiple IQs from large numbers of equivalent tests; nor for that matter do we have direct knowledge of the SEM. All we typically possess is a reliability coefficient (e.g., a test–retest correlation from normative studies) plus one obtained score from a single test administration. How can we possibly use this information to determine the likely accuracy of our obtained score?

Computing the Standard Error of Measurement

We have noted several times in this chapter that reliability and measurement error are intertwined concepts, with low reliability signifying high measurement error, and vice versa. It should not surprise the reader, then, that the SEM can be computed indirectly from the reliability coefficient. The formula is

$$\text{SEM} = \text{SD}\sqrt{1 - r}$$

where SD is the standard deviation of the test scores and r is the reliability coefficient, both derived from a normative sample or other large and representative group of subjects.

We can use WAIS-R Full Scale IQ to illustrate the computation of the SEM. The SD of WAIS-R scores is known to be about 15, and the reliability coefficient is .97 (**Wechsler, 1981** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1732)). The SEM for Full Scale IQ is, therefore,

$$\text{SEM} = 15\sqrt{1 - .97}$$

which works out to be about 2.5.

The SEM and Individual Test Scores

Let us consider carefully what the SEM tells us about individual test results, once again using WAIS-R IQs to illustrate a general point. What we would really like to know is the likely accuracy of IQ. Let us say we have an individual examinee who obtains a score of 90, and let us assume that the test was administered in competent fashion. Nonetheless, is the obtained IQ score likely to be accurate?

In order to answer this question, we need to rephrase it. In the jargon of classical test theory, questions of accuracy really involve comparisons between obtained scores and true scores. Specifically, when we inquire whether an IQ score is accurate, we are really asking: How close is the obtained score to the true score?

The answer to this question may seem perturbing at first glance. It turns out that, in the individual case, we can never know precisely how close the obtained score is to the true score! The best we can do is provide a probabilistic statement based on our knowledge that the hypothetical obtained scores for a single examinee would be normally distributed with a standard deviation equal to the SEM. Based on this premise, we know that the obtained score is accurate to within plus or minus 2 SEMs in 95 percent of the cases. In other words, Full Scale IQ is 95 percent certain to be accurate within ±5 IQ points. This range of plus or minus 5 IQ points corresponds to the 95 percent **confidence interval** for WAIS-R Full Scale IQ because we can be 95 percent confident that the true score is contained within it.

Testers would do well to report test scores in terms of a confidence interval because this practice would help place scores in proper perspective (**Sattler, 1988** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1437)). An examinee who obtains an IQ of 90 should be described as follows: "Mr. Doe obtained a Full Scale IQ of 90 which is accurate to ±5 points with 95 percent confidence." This wording helps forewarn others that test scores always incorporate some degree of measurement error.

The SEM and Differences between Scores

Testers are often expected to surmise whether an examinee has scored significantly higher in one ability area than another. For example, it is usually germane to report whether an examinee is stronger at verbal or performance tasks or to say that no real difference exists in these two skill areas. The issue is not entirely academic. An examinee who has a relative superiority in performance intelligence might be counseled to pursue practical, hands-on careers. In contrast, a strength in verbal intelligence might result in a recommendation to pursue academic interests. How is an examiner to determine whether one test score is significantly better than another?

Keep in mind that every test score incorporates measurement error. It is therefore possible for an examinee to obtain a verbal score higher than his or her performance score when the underlying true scores—if only we could know them—would reveal no difference or even the opposite pattern! (see **Figure 3.14** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/ch03lev1sec18#ch03fig14)). The important lesson here is that when each of two obtained scores reflects measurement error, the difference between these scores is quite volatile and must not be overinterpreted.

The **standard error of the difference** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm01#bm01gloss305) between two scores is a statistical measure that can help a test user determine whether a difference between scores is significant. The standard error of the difference between two scores can be computed from the SEMs of the individual tests by the following formula:

$$SE_{diff} = \sqrt{(SEM_1)^2 + (SEM_2)^2}$$

where SE_{diff} is the standard error of the difference and SEM_1 and SEM_2 are the respective standard errors of measurement.

It is assumed that the two scores are on the same scale or have been converted to the same scale. That is, the tests must have the same overall mean and standard deviation in the normative sample. By substituting $SD\sqrt{1 - r_{11}}$ for SEM_1 and $SD\sqrt{1 - r_{22}}$ for SEM_2, we arrive at

$$SE_{diff} = SD\sqrt{2 - r_{11} - r_{22}}$$

We return to our original question to illustrate the computation and use of SE_{diff}. How is an examiner to determine whether one test score is significantly better than another? In particular, suppose an examinee obtains Verbal IQ 112 and Performance IQ 105 on the WAIS-R. Is 7 IQ points a significant difference?

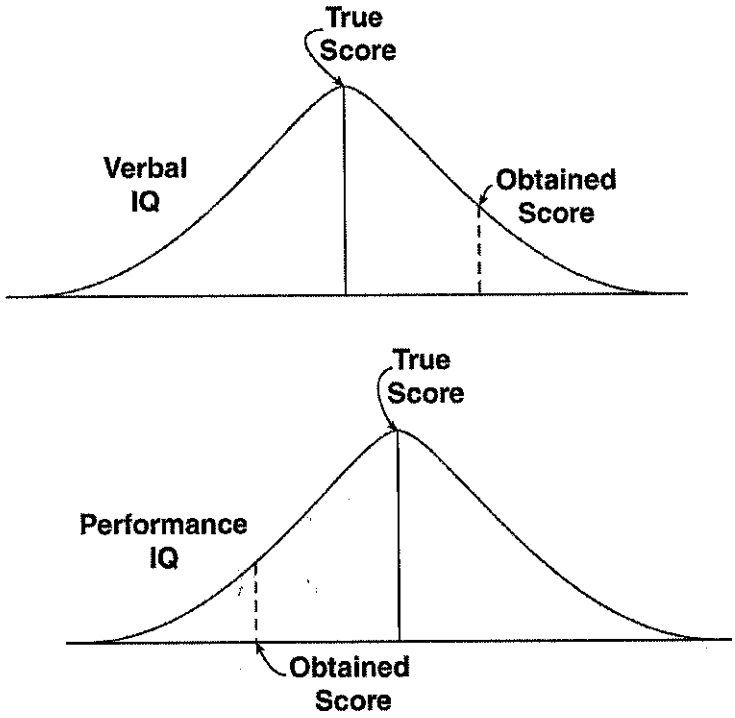


FIGURE 3.14 Obtained Scores Reflect Measurement Error and May Obscure the Relationship between True Scores

Note: In this hypothetical case the obtained Verbal IQ is higher than the obtained Performance IQ, whereas the underlying true scores show the opposite pattern.

We know from the WAIS-R Manual (**Wechsler, 1981** (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1732)) that Verbal and Performance IQ each have standard deviations of approximately 15; and their respective reliabilities are .97 and .93. The standard error of the difference between these two scores can be found from

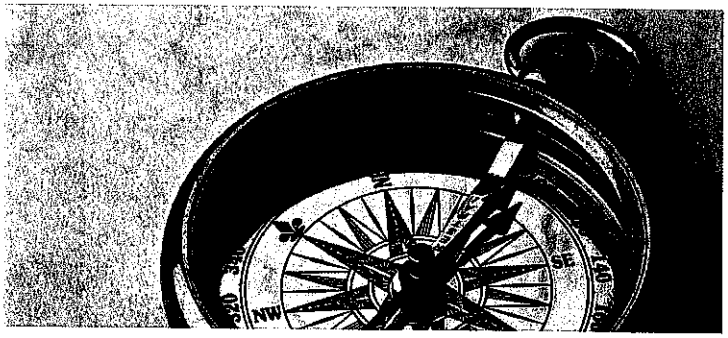
$$SE_{diff} = 15\sqrt{2 - .97 - .93} = 4.74$$

Recall from the discussion of normal distributions that 5 percent of the cases occur in the tails, beyond ±1.96 standard deviations. Thus, differences that are approximately twice as large as SE_{diff} (i.e., 1.96 × 4.74) can be considered significant in the sense that they will occur by chance only 5 percent of the time. We may conclude, then, that differences of about 9 points or more between Verbal and Performance IQ likely reflect real differences in scores rather than chance contributions from errors of measurement. Thus, more likely than not, a difference of merely 7 IQ points does not signify a bona fide, significant difference between verbal and performance intelligence.

[2]This would hold true for subjects of similar age. The SEM may differ from one age group to the next—see Wechsler (2008 (http://content.thuzelearning.com/books/Gregory.8055.17.1/sections/bm02#bm02bib1737)) for an illustration with the WAIS-IV.

REFERENCE & American Psychological Association (2010). Retrieved From http://www.apa.org/ethics/code/index.aspx?item=12

Ethical Principles of Psychologists and Code of Conduct



Including 2010 and 2016 Amendments

- **Introduction and Applicability**
- **Preamble**
- **General Principles**
- **Section 1: Resolving Ethical Issues**
- **Section 2: Competence**
- **Section 3: Human Relations**
- **Section 4: Privacy and Confidentiality**
- **Section 5: Advertising and Other Public Statements**
- **Section 6: Record Keeping and Fees**
- **Section 7: Education and Training**
- **Section 8: Research and Publication**
- ▼ **Section 9: Assessment**

9.01 Bases for Assessments

(a) Psychologists base the opinions contained in their recommendations, reports, and diagnostic or evaluative statements, including for testimony, on information and techniques sufficient to substantiate their findings. (See also Standard 2.04, Bases for Scientific and Professional Judgments (?item=5#204) .)

(b) Except as noted in 9.01c (#901c) , psychologists provide opinions of the psychological characteristics of individuals only after they have conducted an examination of the individuals adequate to support their statements or conclusions. When, despite reasonable efforts, such an examination is not practical, psychologists document the efforts they made and the result of those efforts, clarify the probable impact of their limited information on the reliability and validity of their opinions, and appropriately limit the nature and extent of their conclusions or recommendations. (See also Standards 2.01, Boundaries of Competence (?item=5#201) , and 9.06, Interpreting Assessment Results (#906) .)

(c) When psychologists conduct a record review or provide consultation or supervision and an individual examination is not warranted or necessary for the opinion, psychologists explain this and the sources of information on which they based their conclusions and recommendations.

9.02 Use of Assessments

(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.

9.03 Informed Consent in Assessments

(a) Psychologists obtain informed consent for assessments, evaluations, or diagnostic services, as described in Standard 3.10, Informed Consent, except 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 (e.g., when participants voluntarily agree to assessment when applying for a job); or (3) one purpose of the testing is to evaluate decisional capacity. Informed consent includes an explanation of the nature and purpose of the assessment, fees, involvement of third parties, and limits of confidentiality and sufficient opportunity for the client/patient to ask questions and receive answers.

(b) Psychologists inform persons with questionable capacity to consent or for whom testing is mandated by law or governmental regulations about the nature and purpose of the proposed assessment services, using language that is reasonably understandable to the person being assessed.

(c) Psychologists using the services of an interpreter obtain informed consent from the client/patient to use that interpreter, ensure that confidentiality of test results and test security are maintained, and include in their recommendations, reports, and diagnostic or evaluative statements, including forensic testimony, discussion of any limitations on the data obtained. (See also Standards 2.05, Delegation of Work to Others (?item=5#205) ; 4.01, Maintaining Confidentiality (?item=7#401) ; 9.01, Bases for Assessments (#901) ; 9.06, Interpreting Assessment Results (#906) ; and 9.07, Assessment by Unqualified Persons (#907) .)

9.04 Release of Test Data

(a) The term *test data* refers to raw and scaled scores, client/patient responses to test questions or stimuli, and psychologists' notes and recordings concerning client/patient statements and behavior during an examination. Those portions of test materials that include client/patient responses are included in the definition of *test data*. Pursuant to a client/patient release, psychologists provide test data to the client/patient or other persons identified in the release. Psychologists may refrain from releasing test data to protect a client/patient or others from substantial harm or misuse or misrepresentation of the data or the test, recognizing that in many instances release of confidential information under these circumstances is regulated by law. (See also Standard 9.11, Maintaining Test Security (#911) .)

(b) In the absence of a client/patient release, psychologists provide test data only as required by law or court order.

9.05 Test Construction

Psychologists who develop tests and other assessment techniques use appropriate psychometric procedures and current scientific or professional knowledge for test design, standardization, validation, reduction or elimination of bias, and recommendations for use.

9.06 Interpreting Assessment Results

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. (See also Standards 2.01b and c, Boundaries of Competence (?item=5#201b) , and 3.01, Unfair Discrimination (?item=6#301) .)

9.07 Assessment by Unqualified Persons

Psychologists do not promote the use of psychological assessment techniques by unqualified persons, except when such use is conducted for training purposes with appropriate supervision. (See also Standard 2.05, Delegation of Work to Others (?item=5#205) .)

9.08 Obsolete Tests and Outdated Test Results

(a) Psychologists do not base their assessment or intervention decisions or recommendations on data or test results that are outdated for the current purpose.

(b) Psychologists do not base such decisions or recommendations on tests and measures that are obsolete and not useful for the curre purpose.

9.09 Test Scoring and Interpretation Services

(a) Psychologists who offer assessment or scoring services to other professionals accurately describe the purpose, norms, validity, reliability, and applications of the procedures and any special qualifications applicable to their use.

(b) Psychologists select scoring and interpretation services (including automated services) on the basis of evidence of the validity of the program and procedures as well as on other appropriate considerations. (See also Standard 2.01b and c, Boundaries of Competence (?item=5#201b) .)

(c) Psychologists retain responsibility for the appropriate application, interpretation, and use of assessment instruments, whether they score and interpret such tests themselves or use automated or other services.

9.10 Explaining Assessment Results

Regardless of whether the scoring and interpretation are done by psychologists, by employees or assistants, or by automated or other outside services, psychologists take reasonable steps to ensure that explanations of results are given to the individual or designated representative unless the nature of the relationship precludes provision of an explanation of results (such as in some organizational consulting, preemployment or security screenings, and forensic evaluations), and this fact has been clearly explained to the person being assessed in advance.

9.11 Maintaining Test Security

The term *test materials* refers to manuals, instruments, protocols, and test questions or stimuli and does not include *test data* as defined in Standard 9.04, Release of Test Data (#904) . Psychologists make reasonable efforts to maintain the integrity and security of test materials and other assessment techniques consistent with law and contractual obligations, and in a manner that permits adherence to this Ethics Code.

Section 10: Therapy

History and Effective Date

Amendments to the 2002 "Ethical Principles of Psychologists and Code of Conduct" in 2010 and 2016

Additional Resources

- 2018 APA Ethics Committee Rules and Procedures (PDF, 197KB)
- Revision of Ethics Code Standard 3.04 (Avoiding Harm)
- APA Ethical Principles of Psychologists and Code of Conduct (2017) (PDF, 272KB)
- 2016 APA Ethics Committee Rules and Procedures
- Revision of Ethical Standard 3.04 of the "Ethical Principles of Psychologists and Code of Conduct" (2002, as Amended 2010) (PDF, 26KB)
- 2010 Amendments to the 2002 "Ethical Principles of Psychologists and Code of Conduct" (PDF, 39KB)
- Compare the 1992 and 2002 Ethics Codes