4

Measuring Student Behavior and Learning

Fredda Brown
Queens College, City University of New York
Martha E. Snell
Curry School of Education, University of Virginia, Emeritus

WHY MEASURE STUDENT BEHAVIOR?

To evaluate the impact of a school program on a student, educators must formulate specific strategies for measurement. Four basic reasons for developing measurement strategies are (a) to document what has occurred, (b) to identify the variables responsible for the occurrence (Zirpoli, 2012), (c) to understand when and why learning is

occurring or not occurring, and (d) to be accountable (Alberto & Troutman, 2012). Measurement strategies enable teachers to better predict future performance, and prediction helps teams decide whether program modifications are necessary. Furthermore, knowledge of measurement strategies allows educators to better understand the findings and implications of published research and determine their relevance to teaching practices (Alberto & Troutman, 2012). Kennedy (2005) points out that human memory is fallible—what people think they see, or remember seeing, is often dramatically different from what actually occurred. Objectively and directly measuring behavior is necessary to make the best decisions regarding our students' education.

Teachers are now required to use strategies that have been demonstrated to be effective (i.e., those founded on scientific evidence). However, using evidence-based strategies is only the first step. Teachers must then be accountable for student outcomes; they must demonstrate either that their students are benefiting from these strategies or that they are making data-based program modifications to try to improve the outcomes.

Using an Evidence Base to Guide Instruction

Being accountable for student outcomes has become increasingly important and has now been explicitly stated in educational policy and law, including the No Child Left Behind (NCLB) Act of 2002 and the Individuals with Disabilities Education Act (IDEA) of 2004. These federal mandates were adopted to ensure the quality of instruction and outcomes for all students with disabilities. One requirement of these regulations is that evidence-based practices be used to guide classroom practice (Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005; Wang & Spillane, 2009). The term *evidence-based practice* refers to educational programs or instructional procedures that have been determined to reliably produce positive student outcomes (Tankersley, Harjusola-Webb, & Landrum, 2008). There has been much discussion in the field concerning what constitutes an evidence basis (Horner et al., 2005; Odom et al., 2005; Tankersley et al., 2008). There are at least two considerations that must be examined in order to understand this issue (Spooner & Brown, 2011): (a) experimental methodologies with adequate strength to identify research meeting quality criteria (Odom et al., 2005; Tankersley et al., 2008) and (b) the number of quality research studies necessary to establish an evidence basis (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005). Although large-group randomized trials are considered to be "the gold standard" of research methodologies, most of the research focusing on students with severe disabilities has used single-subject experimental designs, and researchers in this field regard such designs as a valid methodology (McDonnell & O'Neill, 2003).

There have been several suggestions regarding the number of single-subject studies required to establish an evidence base. For example, Horner et al. (2005) recommends that a practice have a minimum of five studies that (a) meet acceptable methodological criteria; (b) have been published in peer-reviewed journals; (c) have been conducted by at least three different researchers, across at least three different geographical locations; and (d) include a total of at least 20 participants. In addition to rigorous research support, the National Autism Center's National Standards Project (2009) includes criteria related to the values and preferences of parents, care providers, and individuals with autism spectrum disorders (ASDs), including a consideration of (a) the side effects of the treatment, (b) whether the treatment is aligned with the values of family members, and (c) whether the individual with ASD is in agreement with the specific treatment. Interestingly, there is no evidence base for determining criteria for what constitutes an evidence base! It might be generally agreed on, however, that the more successful demonstrations of an educational practice, the more confidence we would have in a true functional relationship between that practice and student performance (Tankersley, Harjusola-Webb, & Landrum, 2008). For strategies that are more contentious or controversial, the criteria for establishing an evidence basis should be even more rigorous.

Accountability Through Evaluation

Using evidence-based strategies is the foundation of good practice. However, just because a teacher uses an evidence-based strategy does not ensure that a given student will benefit from it. Ongoing evaluation of individual student progress is an integral part of the teaching–learning process. With an increasing focus on educating students with severe disabilities alongside their typically developing peers in general education settings, data collection strategies become even more of a challenge. Each teacher must balance the need for data to make instructional decisions and to evaluate program effectiveness with the needs of the regular classroom. Although many teachers question the value of data and find data difficult to manage, it is widely accepted that teachers make better instructional decisions when they found them on student performance data (Farlow & Snell, 2005; Janney & Snell, 2013; Zirpoli, 2012).

Teachers need to be accountable for two levels of measurement: process measures and outcome measures (Haring & Breen, 1989). *Process measures* focus on precise, small units of behavior, such as a student's performance on individual steps in a complex chain (e.g., number of prompts needed to complete a 20-step task analysis for using a vending machine), or a discrete skill (e.g., frequency of positive interactions with peers) (Haring & Breen, 1989). These types of data are important in the evaluation of instructional programs as they guide routine decision-making regarding program modifications. *Outcome measures* do not provide this level of detailed information. Instead, these measures offer information regarding the general effects of a program on a person's quality of life reflecting a range of significant outcomes for the individual, school, family, and community. Outcome measures include, for example, time spent in social interactions with non-disabled peers, or presence in school activities.

Meyer and Evans (1993) included in their criteria for successful outcomes of behavioral interventions those which are related to self-determination and quality of life, such as less restrictive placements, greater participation in integrated school experiences, subjective quality-of-life improvements (e.g., happiness, satisfaction, choices, and control), improvements perceived by significant others, and expanded social relationships and informal support networks. Carr (2007) encapsulates these important quality-of-life outcomes as happiness and personal satisfaction and proposes that these hard-to-define variables must be our vision for the future.

Think about the case of a young man who lived in a facility where he received contingent electric shock for the self-injurious behavior of scratching (Gothelf & Brown, 1998). He wore a shock device 24 hours per day at his school program and in his group home. After many thousands of shocks, his self-injury was significantly reduced although not entirely. Using the process measure of "rate of scratching" as the only measure of success, it could be concluded that the contingent electric shock was effective. However, this man's program was judged as being far from effective, not only because it applied an unacceptable strategy, but also because there was little impact on the quality of his life. Consider the following outcomes that accompanied the behavior reduction: (a) wearing a shock device 24 hours per day, (b) the inability to manage his own behavior without the shock device, (c) going to school and living in a group home with individuals who all had severe behavior problems, (d) having no control over daily activities (e.g., what to eat, when to go to bed, or wake up), (e) limited social interaction with individuals without disabilities, (f) limited social interaction with unpaid individuals, (g) self-reports of being unhappy, (h) no control over his future, and (i) living a great distance away from his family. Outcome measures would reveal just how limited this intervention was for this young man—in terms of both the lack of qualitative life outcomes and the use of a strategy that is in opposition to acceptable practice.

Measurement not only must accurately describe current performance on priority skills, but also must reach beyond this traditional assessment of isolated skill increases

or behavior reduction to assess the outcomes that make a significant difference in the individual's life. In the previous example of the young man who participated in an aversive intervention program, meaningful outcomes would include, for example, self-management of his behavior, living in his own home and near his family, designing his own daily routines, pursuing and acquiring a job of his preference, and selecting his own personal care assistants.

Much of this chapter explains how to describe, measure, and graph student behavior and learning with admittedly more focus on process measures and less on outcome measures. However, it is not sufficient to just measure behavior; measurement must be meaningful. Albert Einstein had a sign in his office that read "Not everything that can be counted counts, and not everything that counts can be counted." So, before we get to the details, we will begin with a discussion of basic foundations and principles for selecting measurement strategies and making measurement meaningful.

FOUNDATIONS OF MEANINGFUL MEASUREMENT

Before reviewing strategies for measuring and graphing student behavior and learning, it is important to understand the nature of meaningful and accurate data. Data, if wisely used, can provide information that is critical to the development, evaluation, and revision of instructional efforts (Farlow & Snell, 1994; Zirpoli, 2012). However, measurements of behavior change are of little importance if they do not provide meaningful information. Three characteristics of student performance data must be present in order for a measurement system to be meaningful: (a) Data must reflect important behavior, (b) data must be contextually appropriate, and (c) data must be sufficiently accurate and reliable.

Measurement of Important Behaviors

Data should reflect important and significant behaviors. Simply being accurate and reliable is not sufficient. For example, a teacher can accurately measure the number of times that Jacob was able to complete a preschool puzzle; however, because a preschool puzzle is not age appropriate for Jacob, who is in the fourth grade, it would not be a meaningful measurement. Teams should consider a series of questions to determine whether their measurement strategies are meaningful:

- Do these data measure behaviors or skills that are valued by the student, his or her parents, and the community or society?
- Do these data reflect the qualitative changes that we hope to see in this student?
- Are the types of changes or the amount of change in the student significant?

Many efforts have been made to describe the criteria for evaluating the validity and importance of behavior changes. Researchers and practitioners have used the following five criteria to evaluate the success of behavioral change efforts: (a) statistical significance, (b) clinical significance, (c) social validity, (d) internal validity, (e) treatment fidelity, and (f) quality of life.

Statistical Significance

Experimental or *statistical significance* involves comparing behavior during or following an intervention with what it was prior to the intervention and asking whether that difference is beyond what might be expected by chance (Kazdin, 1976). Often, statistical analysis is used to evaluate the success of an intervention in a research study, but it is not sufficient as the sole criterion for evaluating change, nor is it practical for teams to apply in school.

Clinical Significance

Therapeutic or *clinical significance* is the importance of the change achieved in the behavior (Kazdin, 1976), or the comparison between the change in behavior that has occurred and the level of change required for the individual to more adequately function in society (Risley, 1970). In other words, if the result of an intervention makes no improvement in the student's life (e.g., enabling him or her to eat independently), even though there is statistical significance, the change does not meet the criterion of clinical significance.

Social Validity

Social validity also refers to the significance of a change in an individual's life. In an analysis of the development of applied behavior analysis, Baer, Wolf, and Risley (1987) stated, "We may have taught many social skills without examining whether they actually furthered the subject's social life; many courtesy skills without examining whether anyone actually noticed or cared; many safety skills without examining whether the subject was actually safer thereafter; many language skills without measuring whether the subject actually used them to interact differently than before; many on-task skills without measuring the actual value of those tasks; and, in general, many survival skills without examining the subject's actual subsequent survival" (p. 322).

Social validity is a concept that addresses qualitative aspects of the educational program. It focuses on the acceptability of the educational goals, the appropriateness of the procedures used to influence behavior, and the social importance of the behavior change (Cooper, Heron, & Heward, 2007; Kazdin, 1977; Wolf, 1978). Social validation procedures can be used to determine whether the learned behavior is functional or meaningful (Kazdin, 1980). There are two methods for determining social validity. *Social comparison* contrasts the student's performance with the performance of the student's non-disabled peers. This standard checks against imposing unnecessarily rigorous performance criteria or stopping instruction before the student reaches a socially acceptable level of performance.

> *At Marc's team meeting, the psychologist suggested that a behavioral objective be developed to teach him to sit in his seat and keep his hands in his lap and his feet on the floor for 15 minutes. His general education teacher, Ms. Kwan, questioned the objective, saying "I'm not sure any of the children in the class can sit like that for 15 minutes!" Ms. Wharton, Marc's special education teacher, suggested that they observe Marc during a variety of kindergarten activities and record data on a few typically developing children to determine the range of their attending duration and the "styles of sitting" displayed by the other students. Ms. Kwan agreed that observing peers would help set a more appropriate goal for Mark's paying attention in class.*

The second method used to determine social validity is *subjective evaluation*. In this method, the opinions of significant people, because of their expertise or familiarity with the student, are used to judge the significance of the behavior change. For example, Whalon and Hanline (2008) looked at the social validity of a reading and language intervention for children with autism. In this study, three students with autism, seven and eight years of age, were taught to ask questions about the book that their teacher was reading to a small group, which included their general education peers. Social validity data were collected via interviews of the children with autism, their parents, and their general education peers. Parents were interviewed to determine whether they noticed changes in their child's reading and social behavior, and whether they thought that the intervention was important. Also, videos of the first intervention session and the last session were shown to the parents, and they were asked to describe any changes that they noted in their children's reading and language skills. The children with autism and their general education peers were individually interviewed and were asked about the usefulness of the intervention and whether they experienced any changes in their relationships with each other. The

results of the social validity measures indicated that the children with autism thought that the intervention helped them better understand what they read. Their general education peers reported that they thought that the intervention was helpful, they enjoyed the sessions, and they had more interactions with their peers with autism. Finally, the parents thought that the intervention was helpful for improving both reading and language skills.

Internal Validity

Internal validity refers to the demonstration that changes in behavior are a function of the independent variable and are not the result of other uncontrolled or unknown variables (Cooper, Heron, & Heward, 2007). That is, teachers should feel confident that the behavior change occurred as a function of the educational intervention, and not some other conditions or events that may influence the behavior.

> *Marc's team reached a consensus that although they did not want to set an objective that he sit in his seat for 15 minutes, having him sit in his seat long enough to complete a short activity would be a valuable goal. Ms. Wharton, familiar with the literature on the positive impact of choice on activity participation, suggested that their intervention be based on providing Marc with more choices of activities, in the hopes of increasing the length of time that he sat in his seat. Following intervention, Ms. Wharton was happy to report to the team that the use of more choices seemed to be effective—Marc was sitting for longer periods of time. His mother, however, questioned this finding when she shared with the team that Marc had been troubled by spring allergies, but that he began to take an antihistamine to treat his allergies. Ms. Wharton then wondered whether the change in Marc's in-seat behavior was a function of his new choice program or taking allergy medication.*

Treatment Fidelity

Treatment fidelity, or *procedural reliability,* focuses on the accuracy of the implementation of the procedures. It answers the question, "Did the educational intervention occur as specified in the treatment plan?" This is an important question to ask. Consider trying to analyze why a new intervention is not working. If we are not sure whether the various team members are implementing the strategy in the same way, we cannot determine whether it is the program that is not effective or whether it is just the inconsistent teaching.

Quality of Life

In 1972 Wolf Wolfensberger introduced the term *normalization* to the field of developmental disabilities. This created a paradigm shift in our view of what is important to teach individuals with severe disabilities. He stated that normalization is the "utilization of means which are as culturally normative as possible in order to establish and/or maintain personal behavior and characteristics which are as culturally normative as possible" (p. 28). *Characteristics that are as culturally normative as possible* provided the field with a criterion to reach that focused on the quality of an individual's life. Increasingly educational programs are looking at this criterion in the evaluation of program success. To determine if an individual has an acceptable quality of life, Wolfensberger asks if the individual experiences (a) normal rhythm of the day, (b) normal rhythm of the week, (c) normal rhythm of the year, (d) normal experiences of the life cycle, (e) normal respect, (f) living in a sexual world, (g) normal economic standards, and (h) normal environmental standards.

Meyer and Evans (1989) delineate eight possible outcomes to evaluate the effectiveness of a teaching program:

- improvement in target behavior
- acquisition of alternative skills and positive behaviors
- positive collateral effects and the absence of side effects

- reduced need for and use of medical and crisis management services for the individual or others
- less restrictive placements and greater participation in integrated community experiences
- subjective quality-of-life improvement—happiness, satisfaction, and choices for the individual
- perceptions of improvement by the family and significant others
- expanded social relationships and informal support networks

Certain measurement procedures in community settings promote less-than-normalized lifestyles and interactions, thus detracting from the individual's quality of life. For example, the teaching staff may be so interested in recording the prompting levels necessary to complete a leisure activity that they forget to notice whether anyone is having fun (Brown & Lehr, 1993). Professionals are recognizing that critical components of program evaluation are choice and control over one's life (e.g., Bambara, & Koger, 1996; Bannerman, Sheldon, Sherman, & Harchik, 1990; Brown & Lehr, 1993; DiCarlo & Vagianos, 2009; Kern & Clarke, 2005; Knoster & Kincaid, 2005; Mechling, Gast, & Cronin, 2006).

It is critical that each individual, to the greatest degree possible considering the child's age, have control over the activities in which he or she participates, retain the option to refuse participation, and maintain control over the sequence in which the activities take place and the times at which the activities occur. Opportunities for self-determination are also associated with better postschool outcomes (McGlashing-Johnson, Agran, Sitlington, Cavin, & Wehmeyer, 2003). Interestingly, Agran, Snow, and Swaner (1999) found that although most of the special education professionals whom they surveyed supported the notion that self-determination was an important outcome, they did not necessarily include IEP goals related to self-determination. Inclusion of these goals on the IEP would better ensure the team's commitment to self-determination.

Evaluating the educational impact on quality-of-life factors is important if we are to take seriously our commitment to effecting meaningful change in the student's life. However, measurement of the qualitative components of life is often challenging. For example, limited cognitive and communication skills make it difficult for educators and even families to understand a student's vision of quality of life (Brown, Gothelf, Guess, & Lehr, 1998). Holburn (2002) points out that often the strategies that are used to improve students' quality of life (e.g., person-centered planning) and the collateral outcomes of these strategies (e.g., community inclusion, improved relationships) are difficult to measure but nonetheless necessary.

Thus, recent trends look beyond simple quantitative reports of progress (e.g., acquisition of isolated skills) and see each individual within the context of a meaningful life. Measurement strategies must support the evaluation of these important outcomes.

Measurement That Is Contextually Appropriate

It is critical that a full range of student data be collected and it is equally important that we are unobtrusive in its collection, storage, and use. Always remember that evaluation information is confidential. It should be available only to the student, family members, and the professionals who are directly involved in the student's program. Just as a typical student's report card is never publicly displayed, neither should a student's progress graphs or data records be displayed for others to see. Graphs and data sheets should be organized and stored in record files that are accessible only to teaching staff and to be shared with the educational team.

The methods selected for directly measuring behavior in teaching settings should be as simple and time-saving as possible (e.g., the frequency or event counts for carefully selected and distinct periods or a count of permanent products). Typically,

teachers avoid methods that require extensive observation time (e.g., frequency counts taken across an entire day) and avoid measurement approaches that interfere with teaching (e.g., a clipboard). Teachers must make some compromises in order to obtain the maximum amount of information with the least effort and time commitment.

Minimizing the obtrusiveness, and potential clumsiness, of data collection is a goal for teachers. There are low-tech approaches as well as an increasing number of high-tech approaches to measuring student behavior that are now available to educators. Low-technology approaches include using an index card or "sticky note" to record a student's performance of a task analysis, wrist or key-chain counters, kitchen timers, stopwatches, calculators, or quiet counters to keep frequency counts and task analytic data, or to time the duration of target behaviors.

Permanent product measures (to be discussed in next section) also offer an unobtrusive way to evaluate student performance. For example, once Marc can enter his classroom reliably and safely in his "arrival program," the teacher may simply measure two products that result from Marc's performance of the entire arrival chain: (a) whether his coat or sweater is hung on a hook and (b) whether Marc is participating in the activity that he placed by his photo on the activity choice chart. This measure would be taken 10 minutes after his arrival. Other examples include using measurements of weight gain to evaluate the effectiveness of a program to reduce rumination and vomiting and using ratings of spillage to evaluate the success of a feeding program. Communication logs between home and school may be examined to ascertain a family's perception of student improvement and satisfaction with an instructional program. The list of examples for user-friendly measures depends on the imagination of teachers and family members with regard to alternative ways to evaluate learning.

More recently, there are increasing numbers of apps available for teachers to use on their smart phones or tablets that allow observers to easily record frequency, duration, and interval measures, as well as summarize and analyze data: For example, see the Behavior Tracker Pro by Marz Consulting Inc (**https://itunes.apple.com/us/app/id319708933?mt=8**); Functional Behavior Assessment Wizard by Whizz-Watt Software (**https://itunes.apple.com/us/app/functional-behavior-assessment/id573375887?mt=8**).

Measurement That Is Accurate and Reliable

When a team has decided that a change in a student's behavior is a goal (e.g., increase social interactions, decrease inappropriate verbalizations, extend the use of a communication system to peers), one of the first steps in the process is to define the behavior. A precise description of the behavior is necessary to ensure that the teacher is consistently observing the same behavior, that others are observing the same thing, and that there is continuity of instruction (Alberto & Troutman, 2012). For example, if "improve manners" is a goal, it is unlikely that the student, parents, teachers, teaching assistants, and therapists automatically agree on what "manners" mean. To some, appropriate table manners mean sitting up straight, arms off of the table, napkin in lap, and chewing with one's mouth closed. Others, however, may feel that some of these components are unnecessarily formal. Indeed, for some students, eating with their mouths closed is physically unrealistic.

To prevent ambiguity, an *operational definition* of the target behavior is created; that is, the behavior must be described in a way that is observable and measurable. Agreement on what constitutes a behavior is critical to the development of reliable and valid measurement and evaluation systems. The description of a behavior must be specific enough to allow two or more observers to read the definition and make the same judgment about the occurrence or non-occurrence of the behavior (Alberto & Troutman, 2012). Table 4–1 compares terms that are vague and descriptions that are observable and measurable. These terms represent a sample of the goals that

TABLE 4–1
Vague Versus Observable Descriptions of Behavior

Student	Vague	Observable
Marc	Interacts appropriately with peers	Waves "hi" to peers when he enters the classroom in the morning
	Has improved his grooming	Washes hands before meals
		Asks for assistance to tie his shoes
	Has increased his academic skills	Identifies numbers by pointing to numerals 1 through 10
		Matches letters to pictures
Christine	Has shown an increase in her community participation	Goes to the grocery store with a peer once each week to purchase snacks for drama class
		Uses her Dynavox to request directions and answer questions in the grocery store
	Understands her job responsibilities	Completes a sequence of job tasks that are recorded on the communication device
		Requests help when needed during job tasks

were determined for Marc and Christine. The concepts of "functional" and "meaningful" behaviors should not be confused with the standard of describing behaviors in observable and measurable ways. Designing an operational definition to objectively describe a behavior does not ensure that the definition is functional or meaningful to an individual. For example, the statement "When shown either a red or blue block, Christine will point to the red or blue block placed in front of her" is observable and measurable. It is not, however, a meaningful activity for Christine, as it is neither age appropriate nor functional.

Watch "Defining a Behavior" at www.youtube.com/watch?v=gdQTlftjjHo.

If data are not accurate and reliable, it would be impossible to make confident judgments or decisions about a student's progress on his or her goals and objectives. Later in this chapter, we will discuss interrater reliability (i.e., whether the target behavior is being recorded accurately) and procedural reliability (i.e., the degree to which the instructional procedures are implemented accurately).

QUANTITATIVE MEASURES

Teams must know whether each student's instructional program is effective in helping the student achieve his or her IEP objectives. Many teachers resist data collection because they feel as if they cannot afford the time. However, such teachers may find four months into the school year that their intervention is not working. Students cannot afford to participate in ineffective interventions for four months! Frequent and ongoing data collection provides ongoing feedback about the student's progress and reveals important information that can guide program improvement.

Rationale

Instructional decisions are enhanced by the use of data and such decisions positively influence student performance (Farlow & Snell, 1994; Kennedy, 2005; O'Neill, McDonnell, Billingsley, & Jenson, 2011). It has been many years since Fuchs and Fuchs (1986) demonstrated that teachers were more effective when they used student performance data instead of subjective judgment when making instructional

decisions, and also found greater improvement in student performance when the teachers used graphed data instead of ungraphed data to make decisions. In another classic study, Utley et al. (1987) established that teachers and teachers in training made more accurate judgments about student performance when they used data (graphed, ungraphed, or both) than when they based their judgments on observation. Yet many teachers continue to resist this important aspect of teaching.

When teachers begin to measure individual student performance, other advantages become obvious. Precise measurement of behavior allows teams to see even small changes in the behavior, giving everyone the message "Keep up the good work!" This encourages the continued use of promising instructional programs. Continuing an instructional strategy is frustrating when one does not feel that any progress is being made. Student performance data allow the team to determine when a program is not as effective as planned so that they can design modifications and not let precious instructional time be wasted.

Student performance data can enhance communication with others in the same ways as precise definitions of behavior. Saying, on the basis of intuition, that someone is "doing better in cooking" or "seems to be initiating interactions more frequently" is vague, subjective, and possibly inaccurate. Making a statement such as "Christine uses her communication board to greet the cashier with only two verbal prompts" or "Jacob now waves hello when he enters the classroom in the morning four out of five days each week" communicates clearer and more objective messages.

Watch "Data Collection—Autism Therapy Video" at www.youtube.com/watch?v=B_xPqWCoHuM.

Measurement Strategies

In this section, we review several ways to measure student performance. The strategy selected should suit the behavior to be measured and the situation. Some of the strategies are easy to use and require little time away from the usual routine; other strategies, however, take more planning and time. These more complex strategies may be used when a challenging situation in the classroom warrants additional measurement precision. Table 4–2 summarizes these measurement strategies: (a) permanent products, (b) frequency recording, (c) percentage, (d) rate, (e) duration, (f) task analytic measurement, and (g) interval recording and time sampling.

Permanent Products

Many behaviors have a concrete result, or product, that lasts. Unlike behaviors that must be directly observed when they occur, behaviors that result in a product or physical outcome can be evaluated after the individual has performed the behavior. For example, Marc's parents need not sit by his bed all night long to observe toileting accidents. Instead, evidence of accidents can be observed by looking at or touching the child's sheet in the morning. Jacob's teacher can simply look at his computer journal to check for the number and length of sentences written. Permanent product measures provide opportunities to detect error and quality patterns (e.g., particular words or sounds that may be problematic for Jacob). Because measurement of permanent products does not require continuous observation, it is convenient for classroom use.

Frequency Recording

Some behaviors are transitory and must be measured as they occur. Frequency recording measures the number of times that a behavior (appropriate or inappropriate) occurs within a specified period (e.g., the number of times that a student throws his or her work materials onto the floor during a 30-minute work session, the number of times that the student greets people appropriately throughout the school day). Frequency has been used to measure many types of behaviors. For example, Cosbey and Johnston (2006) used a naturalistic approach to teach three children with severe multiple disabilities to use a voice output communication aid (VOCA) to request access to

TABLE 4–2
Measurement Procedures That Are Appropriate for Classrooms

Description of Measurement	Advantages	Disadvantages	Examples of Behaviors Measured
Permanent products Direct measurement of the lasting and concrete results of a target behavior	• Does not require continuous observation • Permits analysis of products for error patterns	• Must have a tangible result • No immediate feedback	• Appropriate behaviors (e.g., the number of newsletters folded and stapled) • Inappropriate behaviors (e.g., the number of buttons ripped from clothing)
Frequency recording The number of times a behavior occurs within a specified period	• Is useful with a wide variety of discrete behaviors • Can be easily accomplished in the classroom • May be converted to a rate	• Necessitates continuous attention during the observation period • Yields less accurate results with high-rate behaviors or behaviors of varying duration • Inappropriate for behaviors of long duration	• Appropriate behaviors (e.g., spontaneous requests for materials needed, initiation of greetings to peers) • Inappropriate behaviors (e.g., talking out, hitting)
Percentage The number correct (or instances occurred) divided by the number of opportunities (or total intervals)	• Useful when the number of opportunities varies across sessions • Can be used to report task analytic measurements, duration, interval, and time-sampling data • Common measure used in academic settings	• Cannot distinguish the number of opportunities from the score • Cannot be used if there is no ceiling on the number of opportunities	• Appropriate behaviors (e.g., independent eating, correct math problems) • Inappropriate behaviors (e.g., hitting, self-injury, cursing, etc., if measured through interval recording or time sampling)
Rate The frequency of a behavior and its relationship to time expressed as a ratio	• Useful when the number of opportunities varies across sessions • Reflects proficiency	• Cannot determine the total time of the observation period	• Appropriate behaviors (e.g., vocational tasks completed per minute, social interactions per hour) • Inappropriate behaviors (e.g., callouts per class period)
Duration The total amount of time in which a targeted behavior occurs in a specified observation	• Yields a precise record of the length of the occurrence of a behavior • May be used to record the duration of each occurrence of a behavior	• Necessitates continuous attention during the observation period • Requires a device to measure the elapsed time (e.g., stopwatch, smart phone app) for optimum accuracy • Inappropriate for frequent behaviors of short duration	• Appropriate behaviors (e.g., attending to lesson, completion of hygiene routine) • Inappropriate behaviors (e.g., tantrums, stereotypical behavior, out-of-seat behavior)
Task analytic measurement A record of the performance of each step in a sequence of behaviors that make up a task	• Useful for most skills in the domestic, vocational, leisure, and community domains • May be used to guide instruction • Enables a measurement of each behavior that makes up a skill • Can be summarized as a percentage or as number of steps	• Requires a good task analysis of the skill being measured • Not suitable for measuring inappropriate behaviors • May focus too much on motor skills, neglecting the qualitative aspects of the task	• Appropriate behaviors (e.g., making the bed, playing a CD, assembly tasks, preparing a snack)

(*continued*)

TABLE 4–2
Measurement Procedures That Are Appropriate for Classrooms (*continued*)

Description of Measurement	Advantages	Disadvantages	Examples of Behaviors Measured
Interval recording A record of the occurrence of behavior within each of the time intervals within a single observation	• Requires less effort than continuous frequency or duration methods • Does not require as precise a definition of a unit of behavior • Applicable to a wide range of behaviors	• Provides an estimate only • The size of the interval must be appropriate for the behavior frequency • Accuracy is facilitated by timers or other time-based devices	• Appropriate and inappropriate behaviors (i.e., any of the behaviors listed for frequency or duration)
Whole interval Records whether behavior occurred *continuously* throughout each of the intervals within the observation session	• Useful when it is important to know that the behavior continues without interruption	• Underestimates the magnitude of the target behavior	• Appropriate behaviors (e.g., on-task behaviors, engagement in play) • Inappropriate behaviors (e.g., out-of-seat behavior, tantrums)
Partial interval Records whether behavior occurred at any time within the interval	• Useful for behaviors that may occur for fleeting moments • Applicable to behaviors of longer duration • Applicable to high-frequency behaviors	• Overestimates the magnitude of the target behavior	• Appropriate behaviors (e.g., social interactions) • Inappropriate behaviors (e.g., out-of-seat behavior, hitting, hand biting, tantrums)
Momentary time sampling Records whether behavior occurred at the moment each of the intervals ends	• Useful for behaviors that tend to persist for a while • Does not require continuous observation • Can be used with more than one student at a time	• Must sample at frequent and relatively short intervals	• Appropriate behaviors (e.g., on-task behaviors, engagement in play) • Inappropriate behaviors (e.g., tantrums, off-task behaviors)

preferred items and/or peers during play activities. The researchers recorded the frequency of unprompted responses (i.e., the student independently activated the VOCA in the presence of peers), as well as prompted and generalized responses.

In order to have meaningful frequency data, it is necessary to specify the length of time and to compare data from the same length of time only. For example, a teacher may report that her student bit his hand 15 times on Monday but only 5 times on Tuesday. This certainly sounds like excellent progress. However, if the teacher observed the student for 3 hours on Monday but only for 1 hour on Tuesday, it is not possible to conclude whether there was any progress.

Behaviors measured in this way should be readily dividable into discrete units, with a clear beginning and an end, and should be easily visible and countable. For example, stereotypical behavior, such as hand waving, may occur at such a high rate that it is impossible to count each instance accurately. Attempting a frequency count of vocalizations may also be difficult if each vocalization does not have a clear beginning and a clear end. For these two examples, another measurement method (such as one of the interval recording measures) should be selected in place of frequency.

Finally, behaviors measured in this way should be relatively uniform in length and not occur for long periods. For example, a parent may report that her child sucked his thumb only two times. This is not helpful information if each occurrence of thumb sucking lasts 45 minutes! The frequency, in this case, does not reflect the amount of behavior. Other measures, such as duration, would be more appropriate for such behaviors. Frequency recordings would accurately measure a student's correct coin and value identification or the number of times that a student greets his teacher and peers.

Percentage

A percentage score can be used when a behavior can occur a fixed number of times in an observation session instead of an undetermined number of times. Percentage is calculated by dividing the number of behaviors observed by the number of opportunities to perform that behavior. This type of measure is used frequently in the general education system to evaluate mastery of academic concepts (e.g., percentage of words correctly spelled, percentage of math problems correctly completed). Marc's teacher used a percentage to measure social interactions when she counted how many times he passed the materials to his neighbor when it was his turn; on one day, she recorded correct responses for two out of the five opportunities occurring during the morning song, or 40% of the opportunities. Percentages can also be used to measure the number of intervals in which a behavior occurred (Christine was on task with her book checkouts, for 8 out of 10 intervals, or 80%, of the time).

Many educators use percentages to measure performance on task-analyzed activities. For example, Mechling and Stephens (2009) taught four young adults (ages 19 to 22 years) to independently complete task-analyzed (non-microwave) cooking tasks (e.g., preparing hot chocolate, chocolate pudding, ravioli, tuna, french fries) using self-prompting strategies. These researchers compared the use of static picture recipes and video prompting strategies, measuring performance by the percentage of task-analyzed steps completed independently. The results demonstrated that all of the students made substantial gains with both the static picture and the video prompting interventions, although the video prompting resulted in the most gains.

Percentage measures are not appropriate when the number of opportunities to perform a behavior is not fixed or controlled. For example, it is inappropriate to write an objective that states that Christine will "greet her peers 80% of the time" if her teacher cannot determine the number of opportunities that she has to greet her peers.

Rate

Rate can be used to determine the frequency of a behavior and its relationship to time. A rate is expressed by the ratio of the number of behaviors divided by the unit of time (e.g., Jacob got out of his seat three times during the 15-minute writing task, or 0.2 times per minute). In vocational training situations, for example, a goal may be to increase the number of cleaning tasks completed in a certain amount of time (e.g., from washing 5 windows in 30 minutes, or 0.17 per minute, to 10 windows in 30 minutes, or 0.33 per minute).

Rate is also a helpful measure when the observation time of a session varies (e.g., to measure the number of spoonfuls of food that Christine eats per minute with her self-feeder when the length of the lunchtime varies). Using an adaptation of prelinguistic milieu teaching, Brady and Bashinski (2008) increased the intentional communication skills (i.e., gestures and vocalizations) of nine students with deaf-blindness. These researchers measured the frequency of the communication acts and then, because of variations in the length of the sessions, converted these measures to rate per minute. Another advantage of using rate to measure performance is that rate reflects both accuracy and speed, or fluency of performance, instead of just accuracy.

Duration

A duration recording is used if the focus is the amount of time that an individual is engaged in a specific behavior or activity. Sometimes it is desirable for a person to increase the amount of time engaged in an activity (e.g., attending to homework, brushing teeth, exercising), and sometimes it is desirable for an individual to decrease the amount of time spent in an activity (e.g., watching television, displaying self-injurious behavior). Duration measures the total amount of time in which a targeted behavior occurs within a specified time.

Delano and Snell (2006) evaluated the effects of social stories on the duration of appropriate social engagement (and the frequency) of four social skills in three elementary school-age students with autism. The intervention consisted of individualized social

stories, responding to comprehension questions, and a 10-minute play session with non-disabled peers. The researchers found that all three students increased their duration of social engagement. Harvey, Baker, Horner, and Blackford (2003) used duration of sleep to explore the presence of sleep problems in individuals with intellectual impairments who were living in community settings. They found that while the duration of sleep for the individuals in their sample was similar to those without disabilities, the quality of sleep was different (e.g., waking up in the middle of the night) as a function of the interaction between the level of the disability and the use of medications.

Duration can be recorded in three ways: (a) as total duration, (b) as percentage of time, and (c) by measuring each occurrence. Using the *total duration* method, the teacher records the total amount of time that the individual spent engaged in the behavior during the observation period. For example, Marc's teacher may be interested in measuring the amount of time that he spends in appropriate play with peers during recess. Before starting the duration measurement, the teacher must operationally define appropriate play for Marc, making sure that it is possible to clearly determine the onset and termination of the behavior. The teacher can then measure the behavior using the timer on her smart phone. All that the teacher has to do is to start the timer when Marc begins playing with his peers. As soon as he stops playing (e.g., participates instead in self-stimulatory behavior), the teacher stops the timer. The teacher starts and stops the timer accordingly for the course of the recess period. The amount of time accumulated on the stopwatch at the end of the period reflects the total duration.

A *percentage of time* can be derived by simply dividing the total time engaged in the activity by the length of the playtime. For example, Marc may have played with his peers for a total of 5 minutes during a 15-minute playtime. The following equation represents the process for determining the percentage of time:

$$\frac{\text{Total duration of behavior}}{\text{Length of observation period}} = \frac{5 \text{ minutes}}{15 \text{ minutes}} = 33\%$$

Although the duration measure is simple and accurate (if the behavior is clearly defined), another piece of information makes the duration measure even more informative: the frequency of each occurrence. For example, we know that Marc participated in playing with his peers for 5 minutes (or 33% of recess); however, we do not know whether he played for 5 minutes in a row or if he played for only 30 seconds at a time but kept returning to the play area. Such information may be valuable in determining the type of intervention to use with Marc to increase his interactions during recess. The method of *measuring each occurrence* provides this information, although it is more time consuming than the previous two methods. To measure the occurrences, the teacher would start the timer when Marc started to play with a peer, click the timer off when he stopped playing, and then record the duration on a data sheet. The teacher would then return the timer to zero. When Marc started to play again, the teacher would start the timer and have it continue until he stopped playing. When Marc stopped again, the teacher would record this duration, and so on. (Alternatively, the teacher could record on the stopwatch in an accumulating manner as before, but record the frequency of instances of playing, which might be easier to do). At the end of the observation period (e.g., 15 minutes), the teacher would have a record of total duration (e.g., 5 minutes), as well as a count of the number of times that Marc started and stopped playing (e.g., eight times). In this case, the goal would be to increase the duration of time that Marc spent playing and to decrease the number of times that he got distracted from playing.

Task Analytic Measurement

Task analytic measurement focuses on a student's performance on a sequence, or chain, of behaviors during teaching or during testing. This type of measurement is the most frequently used method of instruction and evaluation of student performance on routines or complex activities. To implement task analytic measurement, a teacher

conducts a task analysis (see Chapter 5), designs a data sheet to record student performance, and then records the student's performance on each of the steps delineated in the task analysis (or some portion of the steps if using backward or forward chaining, or partial participation).

During teaching, there are a variety of ways in which the teacher may record student performance: by recording a plus (+) or minus (–), or by recording the prompt level (e.g., verbal or physical prompt) required for the student to complete the step. If the teacher is using a total task-chaining strategy (see Chapter 5), then all of the steps in the task analysis are scored. If the teacher is using forward chaining or backward chaining or partial participation, then only the steps that the student is working on are scored. Table 4–3 shows a task analysis for the skill of making a peanut butter sandwich. Because Jacob can participate extensively in this type of activity, his teacher, Ms. Fuentes, chose to use a total task-chaining strategy and a least-to-most

TABLE 4–3
Jacob's Task Analysis for Making a Peanut Butter Sandwich

Name: Jacob **Teacher:** Ms. Fuentes

Activity: Making a peanut butter sandwich

Materials: Peanut butter, bread, butter knife, plate, napkin

Record number that indicates amount of assistance: 3 – Independent, 2 – Verbal, 1 – Model, 0 – Teacher completes

	Dates			
Routine Steps	**9/06**	**9/13**	**9/27**	**10/04**
1. Initiate snack by going to resource room.	3	3	3	3
2. Go to refrigerator.	2	2	2	2
3. Get out peanut butter.	2	2	2	2
4. Put peanut butter on counter.	2	2	3	3
5. Get bread from bread box.	1	1	1	1
6. Put bread on counter.	2	2	2	3
7. Get butter knife.	1	1	1	2
8. Get plate.	1	1	1	1
9. Put knife and plate on counter.	3	3	2	3
10. Open bread bag.	1	1	1	2
11. Remove two slices of bread and put on plate.	3	3	3	3
12. Open peanut butter jar.	0	0	0	0
13. Scoop out peanut butter with knife.	1	1	2	1
14. Spread peanut butter on one slice.	3	3	3	3
15. Repeat until preferred thickness.	3	3	3	3
16. Put other slice on top.	1	2	1	2
17. Put knife in sink.	1	1	1	2
18. Put peanut butter away.	2	2	2	2
19. Get napkin.	1	2	2	2
20. Bring sandwich and napkin to table.	3	3	3	3
Total	36/60 (60%)	38/60 (63%)	38/60 (63%)	43/60 (72%)

prompting procedure. But, as Jacob's behavior has clearly indicated that he prefers not to be physically guided or touched, his team decided to teach Jacob using just the verbal and model prompts. It was agreed that if Jacob did not complete a step of the task analysis independently after a verbal prompt or a model, which was typically sufficient, the teacher would simply complete that step for him and allow him the opportunity to perform the next step of the sequence. Jacob's mother said that she would also practice this skill with him when he gets home from school.

In general, Jacob's data reflect that he is slowly moving toward the criterion. At the beginning of the week, he achieved 60%, and at the end of the week, he achieved 72%. Keeping track of the individual steps of his task analysis allows more detailed analysis that can contribute to constructive program modifications. Looking at individual steps on his data sheet reveals a few steps that seem to be problematic for him. For example, Jacob is unable to open the peanut butter jar (Step 12) and consistently requires *models* to complete Step 5 (getting bread from the bread box) and Step 8 (getting a plate). This information will allow her team to consider some modifications or adaptations for these steps. For example, should Jacob be provided with a rubber gripper that might make opening the jar easier? Or should he be taught instead to ask someone for assistance to open the jar? Could the bread box and the plates be moved closer to him to make accessing them easier? Or might pictures be helpful as a prompt to remind Jacob of where the items that are needed are kept?

Table 4–4 shows another task analysis for the same activity, but this time for Christine, who is expected to participate partially in the activity instead of performing

TABLE 4–4
Christine's Task Analysis for Making a Peanut Butter Sandwich

Name: Christine **Teacher:** Ms. Washington

Activity: Making a peanut butter sandwich

Materials: Peanut butter, bread, butter knife, plate, napkin, meal preparation overlay for communication board

Record number that indicates amount of assistance: 5 – Independent, 4 – Verbal, 3 – Model, 2 – Partial physical, 1 – Full physical, X – Teacher or peer completes

Routine Steps	Dates				
	10/07	10/08	10/09	10/10	10/11
1. Initiate snack by activating communication device.	3	3	3	3	4
2. Peer assists her into home economics class.	X	X	X	X	X
3. Press switch when "peanut butter" is scanned on communication device.	2	2	2	3	4
4. Press switch when "bread" is scanned on communication device.	2	2	2	2	3
5. Teacher gets items, puts them on counter.	X	X	X	X	X
6. Press switch when "knife" is scanned on communication device.	2	2	2	2	3
7. Press switch when "plate" is scanned on communication device.	2	2	2	2	3
8. Press switch when "napkin" is scanned on communication device.	2	2	2	2	3
9. Teacher gets items, puts them on counter.	X	X	X	X	X
10. Teacher makes sandwich.	X	X	X	X	X
11. Teacher gives her a sample of sandwich and asks whether sandwich is OK.	X	X	X	X	X
12. Press switch when "yes/no" is scanned on communication device.	4	4	4	5	5
13. Teacher fixes sandwich as necessary.	X	X	X	X	X
14. Press switch when "thank you" is scanned on communication device.	2	2	2	2	2
Total	19/40 48%	19/40 48%	19/40 48%	21/40 53%	27/40 68%

all of the steps. Christine is scored on eight steps of the task and her teacher (or peer for Step 2) completes the steps that are marked with an X. Extension and enrichment skills (e.g., initiation, social skills, monitoring skills) are skills where Christine can likely achieve independence. These components allow for meaningful participation in, and control of, the activity even though she cannot perform most of the motor components of the task (see Chapter 5). The prompt procedure for Christine includes physical assistance.

During testing, task analytic measurement can be carried out by using either a *single-opportunity* or a *multiple-opportunity* method. The easiest, although less informative, method is the single-opportunity method. This approach is carried out as follows:

1. Conditions (including materials) are arranged as planned in the instructional program.
2. The instructional cue (if any) is given when the student is attending.
3. The student's independent response to each step in the task analysis is recorded until an error occurs.
4. The following rules can be used to handle errors, periods of no response, and inappropriate behavior:

   - Testing is stopped after the first error and all remaining steps are scored as errors.
   - After a specified latency period of no response (e.g., 3 seconds), testing is stopped and all remaining steps are scored as errors.
   - After a specified period of inappropriate behavior (e.g., 10 seconds of stereotypical behavior) or after a single inappropriate response (e.g., throwing the soap or the towel), testing is stopped and all remaining steps are scored as errors.

For many tasks, the steps performed are scored as correct if they correspond to the task description, regardless of the order in which they are carried out, as long as the result is satisfactory. For example, it is not important whether Marc pulls his right or left arm out of his coat first. However, for many other tasks (e.g., certain assembly tasks), performing each step in order is crucial to the successful completion of the activity. In tasks where order is important, the first step out of sequence is scored as an error. In addition, when the rate of performance is important (as specified in the criteria or standards), the maximum length of time allowed is specified.

Once each week, Ms. Bowers probed Jacob's performance with regard to packing his book bag. Using the single-opportunity method, Ms. Bowers observed Jacob following two instructional cues. First, she gave the entire class the instruction that they need to get their things together to go to music. Second, she told Jacob, as she always did before activity changes, to check his picture schedule to see what the next class is and what he needs to take with him. Following the natural cue of the other students in the class, he began to gather some things together. Ms. Bowers scored a plus (+) on her data sheet for the first three steps: (a) taking out his book bag, (b) unzipping it, and (c) putting in his pencils. After these first three steps of the task analysis, Jacob stood up to join his friends in the front of the room. Ms. Bowers scored a minus (–) on her data sheet to indicate that he did not independently perform the fourth step of putting his music book into his bag. According to the single-opportunity testing method, all remaining steps were scored with a minus (–). Ms. Bowers then proceeded to implement teaching by giving him a verbal prompt for the missed step.

The single-opportunity method is generally completed quickly. It provides a conservative estimate of the student's skills. Less instructional time is wasted because teaching can begin immediately after the first error. Furthermore, learning is less likely to occur during testing; therefore, the single-opportunity method provides a more accurate estimate of the effect of instruction. However, a disadvantage of the method is that performance on task analytic steps that occur after the first error are

not measured because testing is terminated at this point. Thus, the student's successful performance of later steps will not be observed. If a teacher is using backward chaining (i.e., teaching the last step first), the single-opportunity probe does not reflect any progress until training advances to the earlier steps in the chain. Therefore, in such cases, the multiple-opportunity probe produces more information.

The multiple-opportunity method uses the following steps:

1. Conditions are arranged as planned in the instructional program.
2. The instructional cue (if any) is given when the student is attending.
3. The student's responses to each step in the task analysis are recorded as correct or incorrect (i.e., performed correctly or not performed at all).
4. Whenever an error occurs after a specified period of no response or inappropriate behavior, the step is completed by the teacher. Instead of the probe ending, the student is given an opportunity to engage in the next step in the chain. Thus, performance on every step can be assessed.

With both assessment approaches, feedback is not provided to the student with regard to performance of the targeted skill. The withholding of feedback differentiates between the conditions of testing (which represent the most difficult conditions specified in the objective) and the conditions of teaching (when prompts and reinforcement are available). For some students, non-contingent reinforcement may be made available (i.e., reinforcement for something other than performance of the task or generic praise like "Keep up the hard work!") to hold their interest during assessment.

Interval Recording

To use interval recording, the observer divides an observation session into short, equal intervals, and the occurrence (or non-occurrence) of the behavior within each interval is recorded. Some observers build in a brief period (e.g., 5 seconds) for recording between-observation intervals that can increase accuracy (Alberto & Troutman, 2012). Interval recording is useful for those behaviors that do not have discrete start or stop times and that vary in length, are continuous (i.e., of longer duration), or occur with high frequency (Alberto & Troutman, 2012). Interval recording has been used to measure both appropriate behaviors (e.g., peer social initiations and interactions in inclusive settings) and inappropriate behaviors (e.g., out-of-seat, stereotypical, aggressive) in schools and other community settings.

There are two types of interval recording strategies: whole interval and partial interval. In *whole-interval* strategies, the observer notes whether the given behavior occurred continuously throughout the interval. For example, Cox, Gast, Luscre, and Ayres (2009) used a 10-second whole-interval recording system to measure the impact of weighted vests on the in-seat behaviors of students with autism and severe disabilities; that is, the researchers observed whether in-seat behavior occurred throughout the *entire* 10-second interval. This is also what Christine's teacher used to measure her participation in drama rehearsals.

> *Christine's drama teacher thought that she was losing interest in the play. Lately, rehearsals were quite tedious, often focusing on just one or two students while the other students sat and read or did homework from other classes. In order to assess Christine's interest, her teacher asked Christine's teaching assistant to conduct a whole-interval recording for 5 minutes, once at the beginning of the class, once in the middle of the class, and once toward the end of the class. Each of the 5-minute periods was divided into 30 brief intervals of 10 seconds. The teaching assistant noted whether Christine was engaged throughout each 10-second interval in either watching the play or interacting with others. Christine's teacher found that Christine was mostly inattentive to the play and had few interactions with her peers during these times. The first observation at the beginning of the class revealed that she was engaged for 10 out of 30 intervals (or 33%). In the middle of the class, she was engaged for 5 out of 30 intervals (or about 17%); at the end of the class, it was only 2 out of 30 intervals (or 6%). These data show that*

Christine's engagement decreased as the class period progressed. Christine's teacher, the teaching assistant, and two of her friends decided that they needed to make sure that Christine had something to do during these downtimes.

In *partial-interval* recording, the observer notes whether the behavior occurs *at all* during the interval instead of whether it occurs continuously throughout the interval. Once a behavior is observed and noted on the data sheet, further observation is not required for the remainder of that interval. Exactly how many times the behavior occurs during each interval is not recorded. Thus, interval recording provides an estimate of the occurrence of behavior. Because of this, interval size must be carefully chosen, only limited conclusions can be drawn from the data, and the data must be interpreted cautiously (Alberto & Troutman, 2009).

Marc seems to be by himself more and more during unstructured times of the day (e.g., free play), not interacting with other children. His special education teacher, Ms. Wharton, decides that they will use a partial-interval recording to get a better idea of how much time Marc is spending alone during these times. Ms. Wharton selects a 10-minute period in the middle of the 9:30 a.m. free play and a 10-minute period during the 11:30 a.m. free play. Then she divides each of these observation sessions into 10 equal 1-minute intervals. Ms. Wharton records a plus (+) in the 1-minute box if Marc has any type of interaction with another student during the interval and a minus (–) if there is no interaction.

Figure 4–1 shows that Marc had interactions with other children in 2 of the 10 intervals during early morning free play (or 20%). The data do not tell us, however, if Marc

FIGURE 4–1
Partial Interval Recording Form for Marc's Peer Interactions

Name: Marc Teacher: Ms. Wharton

Date: November 2 Behavior: Peer interaction

Code: (+) peer interaction; (–) no peer interaction

9:30 Free Play

10 minutes

Minutes	1	2	3	4	5	6	7	8	9	10
	–	+	–	–	–	+	–	–	–	–

Total: 20%

11:30 Free Play

10 minutes

Minutes	1	2	3	4	5	6	7	8	9	10
	–	+	+	+	–	+	+	–	+	–

Total: 60%

Comments: *Seemed to initiate and sustain interactions when musical toys are involved. When other children got loud, Marc seemed to move away.*

had two very brief interactions with other children or spent a full 2 minutes interacting with them. Marc had interactions with other children in 6 out of the 10 intervals (60%) during the 11:30 a.m. free play.

If Ms. Wharton repeated this observation for a week, following implementation of an instructional strategy to increase interactions, she may still find no change in the data. However, it is possible that Marc is having a significantly greater number of interactions with his peers, but because the interactions were clustered within two or three of the intervals, the progress cannot be seen. In this case, the results of the interval recording may be misleading.

Data from the 11:30 a.m. free play show a higher percentage of peer interaction (60% of the intervals). Continued recording will reveal either that this is an unusual day or that it is typical for Marc to interact more with his peers during free play. Knowing this, Ms. Wharton can analyze that period to determine what variables might be contributing to the increased level of peer interaction. Comments on the data sheet suggest, for example, the possibility that when other students are loud, their noise may inhibit Marc from interacting with them. Ms. Wharton can now investigate further the role that noise plays in Marc's peer interactions.

Selection of the appropriate interval method should be guided by the characteristics of the behavior and the goals of intervention. If the behavior is brief and the goal is to have the behavior occur on a consistent but not necessarily continuous basis (such as Marc's interactions), the partial-interval method should be used or the behavior will not be detected (that is, it is not likely to occur continuously). Other behaviors, such as attention to a task, appropriately occur in a more continuous manner. Such behaviors are best measured by the whole-interval strategy.

When measuring behaviors targeted for reduction, it is best to use the method that provides the most rigorous information. For example, if you were using the whole-interval method to record the extent of self-injurious behavior, the teacher would check the interval only if the self-injury occurs for the entire length of the interval. This would mean that it was possible for the data to reflect no occurrences when, in fact, the student engaged in extensive self-injury, but not *continuous* self-injury. In this situation, the partial-interval method would be more appropriate.

The length of the interval depends on both the behavior being observed (its average length and frequency) and the observer's ability to record the behavior. Interval length is usually measured in seconds (e.g., 5, 10, 30 seconds). The more frequent the behavior, the smaller the interval for observation should be so that the data yield a more accurate representation of behavior.

If large intervals (e.g., 15 minutes) were used with partial-interval recording to measure Christine's frequent behavior of smiling, for example, 100% would be the typical result. In other words, it is almost certain that Christine would smile at least once during every 15-minute interval. This is not informative because it does not provide information about the density of the behavior (e.g., whether Christine smiled 5 times or 150 times).

For behaviors that occur infrequently, the partial-interval method can have longer intervals. For example, because Jacob infrequently initiates interactions with his peers, observing for 30-second intervals makes no sense. It is not likely that Jacob will display the behavior within 30 seconds. However, if the interval is too large, any instance of the behavior can artificially inflate the percentage of intervals. Thus, if Jacob interacts only two times within an hour but the intervals are 30 minutes long, then, statistically, Jacob interacted for 100% of the intervals! This certainly does not reflect the quality of Jacob's behavior. Five-minute intervals might be more appropriate.

Interval recording cannot be done casually, because a teacher's total attention must be directed toward watching the student and timing the intervals during the entire observation period. The teacher must know when to move from one interval to the

next. It can be challenging to teach and collect interval data at the same time (Alberto & Troutman, 2012). A timer of some type (e.g., stopwatch, iPhone) can be used to time the intervals, but checking the time interrupts the observer's concentration; a portable tape recorder with a tape of prerecorded intervals and earplugs may eliminate this problem, although it is somewhat obvious. Teachers must be sensitive to the environment and should be as unobtrusive as possible. For example, it would be distracting to have a beep sound every 10 seconds when observing Christine attending to her drama class.

Time Sampling

Time sampling is a type of interval measure that can be used more practically in teaching settings. As in the whole- or partial-interval recording strategy, a specified observation period (e.g., 30 minutes) is divided into smaller units (e.g., 5-minute intervals). However, unlike interval recording, where a teacher observes the behavior throughout the entire interval, the teacher observes the student only at the *end* of the interval. Time sampling usually uses longer intervals (minutes) than does interval recording (seconds). The teacher records on the data sheet whether the student was engaging in the target behavior at the end of each interval. For example, Reinhartsen, Garfinkle, and Wolery (2002) investigated the effect of child choice of toys versus teacher choice of toys on the engaged and problem behaviors of three 2-year-old boys with autism in an inclusive preschool classroom. Using a time-sampling strategy, these researchers divided the play period into small intervals of time (30 seconds) and observed the children's behaviors at the end of every 5 seconds. The researchers found that when children chose the toys to play with, there was an increase in engaged time and fewer problem behaviors for two of the three boys.

Implementation of time sampling can be done flexibly. Instead of continuously observing and recording at the end of each interval, the teacher can set up random intervals within an observation period. For example, the teacher may decide on an observation period of 1 hour and preselect six random times to observe (instead of exactly every 10 minutes).

It is also possible to use time sampling throughout the day and pre-identify observation times. For example, Ms. Wharton could choose to use time sampling (instead of interval recording) to record Marc's interactions with peers throughout the 3-hour day at school (e.g., free play, snack, lunch, circle time). This strategy is relevant if the goal is for Marc to increase interactions with his peers across many activities, not just during playtime.

Because time sampling does not require continuous observation, teaching and data collection can occur simultaneously (Alberto & Troutman, 2012). In addition, because the observation is so quick (e.g., whether the behavior was occurring or not occurring at that moment), teachers can use the strategy with more than one student at a time. For example, a teacher could record the on-task behavior of a group of students at the end of every 2-minute interval during independent seat work. Every 2 minutes, the teacher would look up and see which of the students were or were not engaged in their independent work (as previously defined by the teacher). She would record a plus (+) for those who were engaged and a minus (–) for those who were not engaged.

Like interval recording, however, time sampling provides only an estimate of the behavior. In fact, for low-frequency and short-duration behaviors, time sampling is even less accurate than interval recording. The less frequent or briefer a behavior, the shorter the interval must be. (Because a teacher is checking at the end of the interval only, he or she may miss the behavior if it does not occur frequently or if it is of short duration.) Thus, time sampling is most appropriate for measuring behaviors that occur frequently and are of long duration.

ORGANIZING STUDENT PERFORMANCE DATA

At this point, we have introduced you to the importance of measuring student behavior and to a variety of quantitative strategies to objectively record behavior. However, if these data are not organized and used, it will be the same as not collecting any data at all. The following section will provide examples of data sheets on which you can record data and then will instruct you on how to visually display, or graph, the data so that the team can understand each student's progress on their IEP goals and objectives.

Designing Data Sheets

Data sheets allow teachers to systematically record data from their observations (e.g., frequency data, task analytic data, interval data). It is important to record this information in a format that will promote subsequent data analysis. For example, some student performance data can provide information for error analyses (e.g., which steps of the task analysis are consistently missed; see Table 4–3), and most data can be converted into graphs for visual analysis.

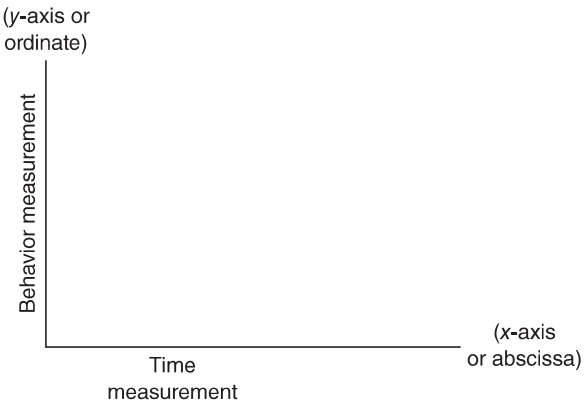
The basic elements of a data sheet are (a) the student's name; (b) the observer's name; (c) the date, time, and location of the observation; (d) the length of the observation; (e) the behavior(s) observed and, if necessary, a brief observable description of each; (f) adequate space for data recording (e.g., room for a 15-step task analysis or ten 2-minute intervals); (g) a scoring code; (h) a data summary; and (i) comments (see Figure 4–1 and Tables 4–3 and 4–4). In addition to providing the range of information necessary to make effective instructional decisions, a data sheet can also assist in functional behavioral assessment (Lohrmann & Brown, 2006) (see Chapter 7 for a discussion of functional behavioral assessment). For example, for the time-sampling procedure used to measure Marc's interactions with his peers throughout the day, his teacher could specify the time of day and the activity in which the data were measured. With this type of information, it is possible to analyze the events or variables (e.g., activities, the time of day, different peers, materials) contributing to the presence or absence of peer interaction. Adding an extra column on the data sheet for recording the incidence of any inappropriate behavior may enable the teacher to see a trend in the relationship between an inappropriate behavior and the time of day or the type of activity.

Graphing Your Data

Analysis of information on a data sheet may provide important details about performance during an instructional session (e.g., specific steps performed correctly or missed on a task analysis, number of interactions at the beginning versus the end of free play). However, significant limitations are encountered if such data are left in this raw form. For example, it is difficult to interpret or analyze behavioral data from a data sheet alone, especially when weeks of data are considered. Behavioral data can be most effectively interpreted and analyzed when they are graphed. Furthermore, graphing data soon after the observation period provides the teacher with immediate feedback regarding performance (Cooper, Heron, & Heward, 2007). Graphs allow teachers to more easily detect trends in a student's progress and thus to make more effective program decisions. When trends are positive, many teachers also find that graphs are reinforcing because they are a continual source of feedback.

Although some teachers initially feel apprehensive or intimidated by graphs, most soon discover that graphs are actually simple to design and read (Hojnoski, Gischlar, & Missall, 2009). A graph is made up of two axes (see Figure 4–2). The *abscissa,* or the *x-axis,* is the horizontal line. The abscissa usually represents the time frame of a measurement (e.g., each data point reflects the data from a day, week, or month). The *ordinate,* or the *y-axis,* is the vertical line. It is labeled with the target behavior

FIGURE 4–2
Basic Components of a Graph



(i.e., dependent variable) being measured (e.g., peer interaction, words read) and the measurement that was used (e.g., duration, frequency, percentage of intervals). For example, on a graph of Marc's interval data for interactions with his peers, his teacher might label the abscissa as "school dates" and the ordinate as "the percentage of 1-minute intervals of interaction with peers."

Converting Data

Before any points can be plotted on the graph, data must be converted into a single numeric form for each data point. Frequency data can be tallied and presented as the total number in a given period (e.g., the number of times that Marc correctly looked at the named peers during the morning group). Total duration data may be presented as the total number of seconds or minutes during which a behavior occurred within a given period (e.g., the number of minutes spent engaged in a 15-minute leisure activity). Duration data that are collected using the method of measuring each occurrence can be presented as a duration (the number of minutes spent interacting with peers during morning free play) or as a frequency (the number of peer interactions during morning free play), or both ways if the teacher thinks that such a presentation would be helpful. Interval or time-sampling data can be converted into the number or percentage of intervals in which the behavior occurred.

Converting task analytic data, which involves multiple steps and may have a range of scoring codes, is a little more complex. There are a number of ways to summarize the data. First, a common approach is to summarize data as the number of steps that the student performed *independently* out of the total number of steps. Then, these data can be converted to the percentage of total steps performed independently. For example, if the student independently performed 3 out of 10 steps on a task analysis, a 30% would be graphed. A second way to summarize the data is to calculate the amount of assistance required to complete the task analysis and convert it into a percentage of independence. For example, in Jacob's 20-step task analysis for making a peanut butter sandwich (see Table 4–3), data are recorded on if he did each step independently (record a 3), or required a verbal prompt (a 2 is recorded), a model (a 1 is recorded), or did not complete the step (a 0 is recorded). The steps for converting Jacob's task analytic data into a single numeric form for graphic presentation are as follows:

1. *Determine the most points that Jacob can earn during each session:* The teacher multiplies the number of steps in the task analysis (20) by the number of points possible in each step (3) for a total of 60 possible points that can be earned.
2. *Add the number of points earned in the session:* Jacob scored a total of 36 points (out of a possible 60) on September 6.

3. *Calculate the percentage:* Divide the number of points earned (36) by the total number of points possible (60) to calculate the performance percentage (in this case, 60%).
4. *Plot the data:* Plot the performance percentage (e.g., 60%) on the graph.

Setting Up a Graph

Once raw data have been converted into a single number to be graphed, it is easy to plot the data point. First, be sure to label the ordinate or vertical line (the *y*-axis) with the behavior being measured and the type of measurement being used (e.g., percentage of independent vacuuming, number of verbalizations during lunch). Next, divide the ordinate into equal intervals that cover the possible range of data (e.g., 0% to 100%, 0 to 50). If there are no definite upper and lower limits, the range should extend from the baseline level to the target level, with some extra space added at both ends to allow for variability.

Data points within each phase of a program (e.g., baseline, intervention, reinforcer change) are connected by straight lines, but they should not be connected across the vertical lines that indicate a phase change (e.g., the change from the baseline to the intervention phase). Each program phase is separated by a broken vertical line and should be labeled at the top of the graph to indicate the intervention used (e.g., baseline, picture prompting, peer model). Data points that represent probe data should be distinct from data points that represent teaching data because the conditions are very different. For example, a teacher might use open data points for the probe data and closed data points for training data. To enhance the effectiveness of a graph for data analysis, date the graph along the abscissa (*x*-axis) using the same time intervals as the data are being recorded (e.g., daily, weekly, monthly). Figure 4–3 lists every week for 3 months because probe data will be recorded for making a peanut butter sandwich on a weekly basis. It is important to delineate the dates before recording to allow automatic skipping of spaces for missing sessions. Because missed sessions can have a detrimental effect on a student's performance, it is important to be able to see these gaps in time. Note that no data are recorded for September 20 in Figure 4–3. Jacob's teacher should investigate the reason for the absence of data collection and the effect of this absence on his performance.

Plotting Data Points

Figure 4–3 shows 4 days of probe data taken from Jacob's data sheet. To plot data, place each data point (i.e., the total for the session) at the intersection of the session date (on the abscissa) and the level of performance (on the ordinate). Note the space between the dates of September 13 and September 27; it is best to skip the space to

FIGURE 4–3
Four Days of Graphed Probe Data for Jacob Making a Peanut Butter Sandwich

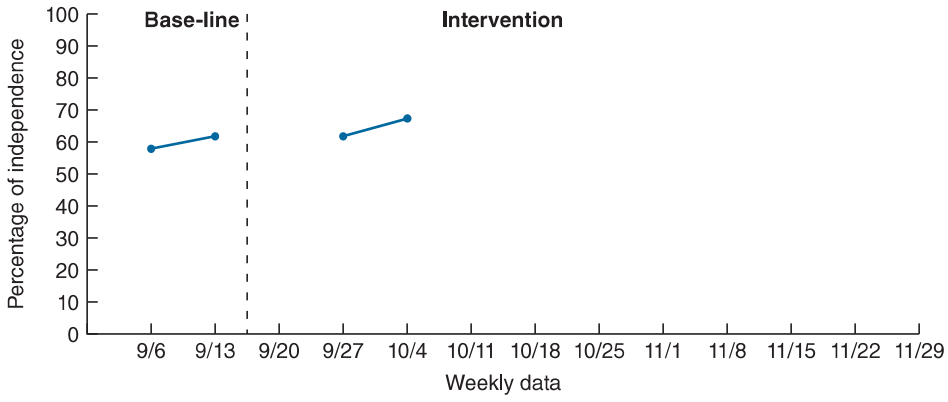
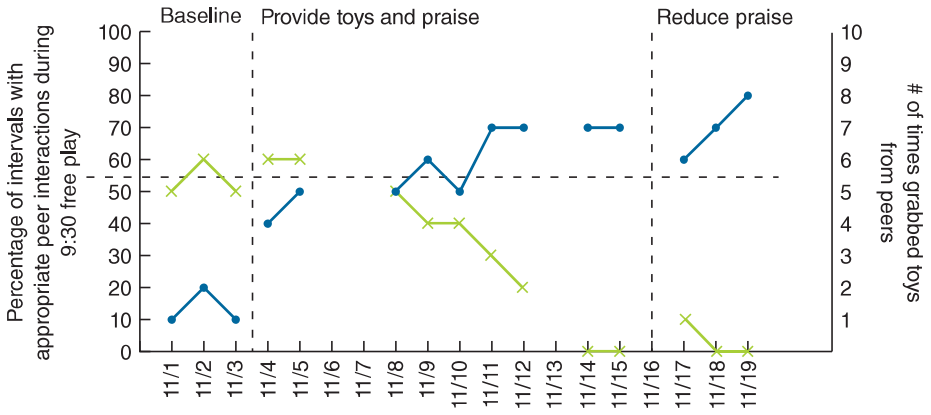


FIGURE 4–4
Use of Graph to Show Marc's Progress and Program Changes During Free Play



indicate the missed session. However, all consecutive sessions in a given phase (according to the scheduled plan of intervention) are connected.

Many teachers find it useful to distinguish between graphs that reflect behavior targeted for acceleration versus graphs that reflect behavior targeted for deceleration. To distinguish between these behaviors, some teachers use a dot to represent acceleration data and an X for deceleration data. This strategy makes successful and unsuccessful trends even more obvious during data analysis. It also allows two related data paths to be plotted on the same graph. Figure 4–4 shows a graph for Marc that uses dots for acceleration data (the percentage of intervals with the appropriate peer interactions during free play) and Xs for deceleration data (the number of times that he grabbed toys from peers). Note that Figure 4–4 uses two ordinates to identify the two targeted behaviors (i.e., peer interactions and the frequency of grabbing) with the two different measurement strategies (i.e., percentage and frequency). Figure 4–4 reflects a successful program. That is, the appropriate behavior (peer interactions) shows an increasing trend and the inappropriate behavior (grabbing) shows a decelerative trend.

Computer-Generated Graphs

Computer-generated graphs have a variety of advantages, including easy storage of data, analysis capabilities, dissemination capabilities (although this can also be a potential danger because of privacy and confidentiality), and professional presentations of data at team meetings. More and more teachers and other professionals are using computers to generate their graphs instead of using a pencil and paper. Hojnoski, Gischlar, and Missall (2009) describe (in a nine-step task analysis) how to use Microsoft® Excel® 1997 to create line graphs. Tutorials for generating computer graphs are readily available online, including how to chart single-subject designs.

Using Self-Graphing Data Sheets

Some teachers find that it is efficient to combine the data recording sheet and the graph. This has been done to measure prompting levels (Alberto & Schofield, 1979), task sequences (Holvoet, Guess, Mulligan, & Brown, 1980), and task analysis (Bellamy, Horner, & Inman, 1979). Plotting data from a task analysis in this manner has been referred to as an upside-down or self-graphing format (Test & Spooner, 1996). Figure 4–5 is a self-graphing data sheet for recording progress on hair brushing.

A self-graphing data sheet can be used to record data by making a slash (/) through the step number if the student responds independently and an X through the step number if the student needs assistance to complete the step. At the end of the

FIGURE 4–5
Self-Graphing Data Sheet for Hair Brushing

Program: Brush hair Student: Jacob Teacher: Ms. Fuentes

Task analysis														
11. Put brush away	11	11	11	11	11	11	11	11	11	11	11	11	11	11
10. Check hair for neatness	10	10	10	10	10	10	10	10	10	10	10	10	10	10
9. Brush left back of head	9	9	9	9	9	9	9	9	9	9	9	9	9	9
8. Brush left side of head	8	8	8	8	8	8	8	8	8	8	8	8	8	8
7. Brush front	7	7	7	7	7	7	7	7	7	7	7	7	7	7
6. Brush right side of head	6	6	6	6	6	6	6	6	6	6	6	6	6	6
5. Brush right back of head	5	5	5	5	5	5	5	5	5	5	5	5	5	5
4. Pick up brush	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3. Select desired materials	3	3	3	3	3	3	3	3	3	3	3	3	3	3
2. Locate brushing materials	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1. Initiate brushing	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	9/2	9/9	9/16	9/23										

Key: ／ = independent
╳ = needs assistance

session, add the number of slashes and then circle the number correct for the day. A graph is formed by connecting the circles over a number of days.

Saving Ungraphed Data

When raw, or ungraphed, data of performance on a multiple-stepped task are summarized and graphed, a certain amount of information is lost. For example, teachers who look at a graph on the accuracy of Jacob's sandwich-making would know about his overall progress on the task across sessions, but they would not know the specific information available on his data sheet, such as which steps he missed or did correctly on a given day or if these missed or correct steps were consistent across days.

Teachers should preserve ungraphed data because response-by-response information may help them make decisions about program implementation when progress is poor. Even with non–task analytic or discrete data (e.g., the number of correct greetings made by Marc at school), graphed summaries lose some of the information that can be preserved on the data collection sheets. For example, Marc's teacher analyzed the ungraphed data to determine whether his performance is better in the early part of the morning or in the later part of his school day, or if he greeted certain peers or adults more frequently than others.

Frequency of Data Collection

Early in the development of current data-based teaching models, it was fairly common to hear the advice that data should be collected each time an instructional activity was implemented. In fact, data collection practices that were recommended for teachers were similar to those used by researchers. Although this may now sound excessive, it was an important phase, as it afforded educators additional understanding of the learning process of students with severe disabilities and increased knowledge of data analysis and evaluation. As educational strategies have become more integrated and community based, data collection procedures have also changed to

better fit these settings (Test & Spooner, 1996). In contexts where teachers of nondisabled peers are not recording data, conspicuous data collection procedures do not enhance an integrated view of a student with severe disabilities. Measures that are most appropriate for inclusive community settings are ones that do not interrupt instruction, take minimal time to complete, are unobtrusive and do not stigmatize the individual, and provide both objective and subjective observations.

Just how often a teacher should take data is a subject of great debate. A teacher may find a wide variety of suggestions when reading the literature (e.g., collect data once each day, every time the skill is practiced, or once a week). For example, a teacher can collect trial-by-trial data or collect probe data, and the teacher may not be the only one collecting data (Alberto & Troutman, 2012). Most discussions of the frequency of data collection, however, suggest that once a skill is fluently performed by the student, data collection could be reduced (Farlow & Snell, 2005).

Although there are still no definitive answers to the question of exactly how much data are needed to make accurate instructional decisions, the following guidelines can assist teachers in deciding how frequently to collect data (Browder, 1991; Brown, 1991; Farlow & Snell, 1994; Snell & Lloyd, 1991):

- Higher priority objectives (i.e., those related to the health and safety of the individual or others) warrant daily data collection. Daily data collection not only is more sensitive to changes in the trend but also contributes to an important functional assessment of the behavior.
- Lower priority objectives, or objectives that are scheduled for instruction on a less-than-daily basis (e.g., grocery shopping), may be evaluated less frequently (e.g., once every one or two weeks).
- Implementation of a new program requires frequent data collection. For the first two weeks of the instructional program, data could be collected daily or at every teaching session if the lessons are not held at least once a day. When the student has shown steady progress (e.g., two weeks of data), data collection could be reduced to weekly.
- Skills being taught to replace problem behaviors can be measured in the same way as any other new program (e.g., initially, data are collected daily; then, following progress, collection is reduced to weekly); however, these probes should be carried out in the context of relevant situations, persons, and environments.
- Data that show progress as planned, with a clear accelerating trend, may be evaluated less frequently, such as on a weekly basis.
- Data that do not show progress as planned or that are variable warrant evaluation on a continual basis, minimally twice a week and ideally on a daily basis.
- Anecdotal records or logs can be used once or twice weekly to record general information concerning a student's overall daily performance and to systematically assess responses to program efforts and any conditions that might affect a student's learning (e.g., tasks or activities that the student enjoys, tasks or activities that the student does not enjoy). Such logs may be useful supplements to more precise, quantified data.

DATA ANALYSIS FOR BETTER DECISION-MAKING

Team members can learn a lot from their graphed data. However, teams need to make sure that the data they collect are accurate. Only if data are accurate can the team be confident that data analysis will help improve the educational program. The remainder of this chapter discusses the importance of accurate data, the use of different types of data, and then provides strategies for analyzing classroom data and ways in which this information can most effectively help educators make instructional decisions.

Measures of Accuracy

Because important decisions are made on the basis of data, team members must have confidence in the data they collect. Consider the following example where the relationship between accurate data collection and effective program evaluation is obvious.

> *Data recorded by Jacob's teaching assistant indicate that he can prepare his snack with only two verbal prompts. However, when his teacher assists Jacob in this activity, she finds that Jacob needs not only significant verbal prompts but also gestural cues. The discrepancy in these data may be because Jacob is not accustomed to preparing his snack with his teacher. If this is the case, then certain programmatic changes can help Jacob generalize his snack preparation skills in the presence of others. However, it is also possible that someone is not recording Jacob's data accurately. If this is the case, changes focusing on generalization would not be appropriate. Efforts should instead focus on increasing the accuracy and reliability of the data collection.*

Interobserver (or Interrater) Reliability

Interobserver reliability is assessed to determine whether the target behavior is being recorded accurately and is the most commonly used indicator of measurement quality in applied behavior analysis (Cooper, Heron, & Heward, 2007). One way to ensure that data are accurate or reliable is to have two independent observers record the behavior of a student at the same time, compare the two observations, and mathematically determine the extent of agreement of the data. The percentage of interobserver reliability can be calculated by dividing the number of agreements between the two observers by the number of agreements plus disagreements and multiplying by 100. The result of this calculation is a *percentage of agreement:*

$$\frac{\text{Agreements}}{\text{Agreements + Disagreements}} = \text{Percentage of Agreements}$$

For example, two teachers use a partial-interval recording to observe the presence of a specified behavior. A 5-minute observation period is divided into ten 30-second intervals. Each time that they observe the target behavior, they record an X in the correct cell. The results of the observation are as follows:

30-second intervals	1	2	3	4	5	6	7	8	9	10
Teacher 1	X	X	—	X	—	X	—	X	X	—
Teacher 2	X	X	X	X	—	X	—	X	X	—

According to this formula, the reliability between the two teachers is as follows:

$$\frac{9 \text{ agreements}}{9 \text{ agreements + 1 disagreement}} = \frac{9}{10} = 90\% \text{ agreement}$$

Generally, a reliability coefficient of 0.80, or 80%, is considered to be acceptable. Poor interrater reliability should prompt the team to improve agreement among its members in ways such as clarifying the behavioral definition of the behavior being observed, offering further training for the staff collecting the data, or simplifying the observational system (Schloss & Smith, 1998).

Procedural Reliability

Procedural reliability (also referred to as treatment integrity, treatment fidelity, and fidelity of intervention) is the degree to which program procedures are implemented accurately. Procedural reliability asks this question: Did the educator follow the instructional plan? If we do not assess the accuracy of the implementation of a

program and a student is not experiencing success, then we could not determine whether it was actually the intervention that was ineffective (Hojnoski, Gischlar, & Missall, 2009; Lane & Beebe-Frankenberger, 2004). Additionally, with the increased number of individuals working with students in inclusive settings (special educator, general educator, paraprofessional), as well as with the varied individuals providing support in inter- and trans-disciplinary service delivery approaches, procedural reliability becomes critical. In an extensive discussion of procedural reliability, Billingsley, White, and Munson (1980) point out that all relevant variables in a program must be evaluated. Program components such as delivery of reinforcers, use of prompts, program setup, antecedent events, and consequent events should be examined. A behavioral checklist for each intervention procedure can be designed, and the teacher can check off each component used (Kerr & Nelson, 2009). Lane and Beebe-Frankenberger (2004) suggest designing checklists for the specific instructional components expected to be implemented by the educator. This checklist would then be used to assess whether the procedure was being followed. The following are some specific questions to consider when assessing procedural reliability:

- Is the instructional plan implemented as frequently as planned?
- Does the instructor use the correct sequence and timing of instructional prompts?
- Does the instructor deliver the appropriate consequences?
- Are instructional cues delivered in the manner designated in the program plan?
- Were all of the necessary instructional materials available?
- Was the program implemented in the correct environment?

Procedural reliability can be calculated in much the same way as interobserver reliability. Billingsley, White, and Munson (1980) offer the following formula:

$$\text{Procedural reliability (\%)} = \frac{(\text{TA} \times 100)}{\text{TT}}$$

In this formula, TA is the number of teacher behaviors in accordance with the program plan, and TT is the total number of teacher behaviors that could have been performed in accordance with the program plan. As an example, Christine was supposed to participate in the library with her nondisabled peers eight times each month (twice each week) but participated only six times last month. Applying this formula to the intervention frequency gives the procedural reliability:

$$\text{Procedural reliability (\%)} = \frac{6 \times 100}{8} = 75\%$$

Teams must feel confident that instruction is having the desired effect on student performance. It is also helpful to know that an intervention is responsible for the change in the student's performance, not just the passage of time or some other event. Sometimes simple, single-subject designs (e.g., reversal, changing criterion) can help teachers feel more confident about the effect of their instruction or the impact of various changes on the educational environment.

Jacob's parents informed his teacher that he was going to be placed on a new medication for a few weeks and that she should be observant to determine whether she noticed any changes in Jacob's on-task behavior. Ms. Fuentes decided to draw a vertical change line on his "on-task" graph to mark the beginning of the medication. At the end of three weeks, Ms. Fuentes was informed that they would be discontinuing the medication; she then drew another vertical line to indicate the discontinuation of the medication. At this point, Ms. Fuentes was able to inform the team that during the time that he was taking the medication, Jacob's on-task behavior improved and that when it was discontinued, the behavior returned to premedication levels. The physician represcribed the medication, and Ms. Fuentes drew another change line and, within a week, saw that the on-task behavior once again increased.

Ms. Fuentes used what is called an ABAB, or reversal design (where *A* is the notation for baseline, and *B* is the notation for intervention), to monitor the changes in Jacob's behavior as a result of the changes in his medication regime made by his family and doctor. After seeing an initial increase in the behavior following the medication, Ms. Fuentes was not confident that it was actually the medication that was responsible for the change in behavior. After all, at around the same time, she added more picture cues to his instruction; perhaps it was the picture cues that helped Jacob stay on task. However, because the behavior decreased *each time* that the medication was withdrawn and increased *each time* that it was prescribed, Ms. Fuentes and the team were more confident that the medication was, at least in part, responsible for the change in Jacob's on-task behavior.

Although an in-depth discussion of single-subject experimental designs (e.g., reversal, multiple baseline, changing criterion) is not possible in this text, we recommend further reading in this area. There are a number of texts that provide comprehensive reviews of single-subject designs (e.g., Alberto & Troutman, 2012; Cooper, Heron, & Heward, 2007; Kennedy, 2005; Kerr & Nelson, 2009; Maag, 2003; Miltenberger, 2008; Zirpoli, 2012). We will limit our later discussion to the simplest classroom design that is non-experimental in nature (i.e., the baseline–intervention, or AB, design).

Types of Data

There are many sources of information that are available to teachers that will help them determine whether students are benefiting from instructional programs. For example, anecdotal records that teachers maintain can be a useful supplement to the quantitative data that is the focus of this chapter. These records may include staff notes on unusually excellent or unusually poor student performance, or comments sent by the teacher to the home, or from the family to the school. Different levels of quantitative evaluation can occur for each objective identified on an IEP. Test conditions and training conditions provide two contexts for obtaining valuable data (Farlow & Snell, 1994).

Probes or Testing Data

Testing or probing means that a person's performance is checked under criterion conditions (i.e., conditions that as closely as possible use natural contexts, cues, and consequences); that is, those conditions under which we ultimately want the behavior to be performed. Thus, the teacher typically provides no prompting or teaching assistance, no reinforcement for task success or improvement, and no corrections. The goal of testing is to learn about a student's current performance under criterion conditions (specified in the objective), not to teach the student. Because students are not being instructed when they are tested, we try to limit the frequency, but testing must be done as frequently as needed to get an accurate picture of the student's performance under criterion conditions.

Teaching Data

In contrast, teaching data can be recorded whenever the student is taught, or as described above, less often in certain circumstances (e.g., low-priority objective). Under teaching conditions, data are recorded while a student is being taught (e.g., least-to-most prompting, reinforcement, error corrections). Learning is the goal of teaching, so conditions are planned to promote improvement in performance and to advance the student through the various stages of learning.

Because of the absence of prompts and reinforcement during probes, a student's performance is typically less proficient than under teaching conditions. Probe data thus represent conservative measures of learning but may more accurately represent a student's performance in natural, unaided situations. Probe data taken for a skill before a teaching program is initiated are called *baseline data*.

When Christine was first learning how to use her communication board to greet the cashier in the grocery store, her teacher initially tested her at the grocery store (i.e., she was given a baseline probe, or test). Probe data showed that Christine was unable to activate the correct greeting symbol. Instruction was implemented to teach her this skill during daily sessions at school in which her teacher recorded her performance during instruction (teaching data). When Christine went to the grocery store, her teacher again recorded probe data to determine how she was performing the skill in the natural context. When Christine performed some, but not all, of the steps, her teacher noted this and then immediately began teaching to prevent any more difficulty.

Obtaining a Baseline

In order to objectively determine whether a student is progressing with regard to a particular objective, we must first know what the student's skills were prior to our intervention. This is done by conducting a *baseline* measurement. The baseline is the period when no intervention or teaching is occurring. Baseline data measure the behavior (the dependent variable) before intervention (the independent variable) is initiated. In other words, the baseline phase of measurement describes a student's performance under the naturally occurring conditions in his or her environment without instructional manipulation.

Ms. Wharton used interval recording to measure Marc's interactions with his classmates during two unstructured periods for three days (Figure 4–1 shows the recording on one day). Because Ms. Wharton observed Marc under natural conditions, without intervening in any way, these three days are considered to be a baseline. After implementation of an intervention program to increase Marc's interactions with his classmates, Ms. Wharton will compare the baseline data with the intervention data to determine whether there was an increase in peer interactions. Ms. Wharton was particularly interested in the effect of the intervention on Marc's behavior during the 9:30 a.m. free play.

A teacher should be cautious about how long a baseline condition is in effect. Generally, the rule is to continue baseline measurement until there is a stable trend in the data. It is considered unethical, however, to continue a baseline measurement in certain situations. First, if a behavior is dangerous, it is unacceptable to wait for a stable trend before beginning treatment. Many times, a teacher can find other forms of data to use as a baseline (e.g., incident reports, daily logs). Second, many students have little or no behavioral activity related to the target objective (e.g., sign language for a student who has never used sign language). Again, it is considered unethical to delay instruction for an extended period of time when a student clearly cannot perform the behavior. Teachers must remember that the baseline does not refer to the absence of a program or to downtime; instead, the baseline refers to the educational context before a given program is implemented or the time when a particular program is withdrawn or stopped. Third, if the direction of the baseline trend is opposite to the direction of the desired trend (e.g., the number of peer interactions are decreasing, or aggression is increasing), collection of baseline data should be discontinued and intervention initiated. When this type of trend occurs, something in the baseline condition is either extinguishing the behavior (e.g., the absence of intermittent teacher praise for playing with others was once reinforced and was withdrawn during the baseline period) or reinforcing the behavior (e.g., a lack of teacher-implemented consequences for aggressive behavior allows the student to get attention from the other students).

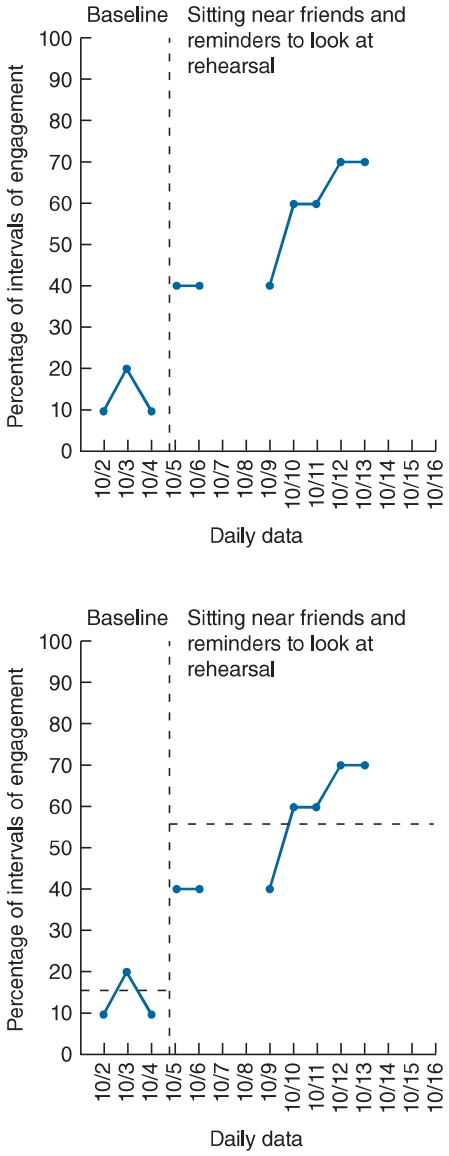
Baseline–Intervention Comparison

Comparing a baseline condition (A) with an intervention condition (B) is called an AB, or baseline–intervention, design. This design is referred to as a *non-experimental* design because no conclusive demonstration of a cause–effect, or functional, relationship between the intervention and the observed changes in behavior is possible. These

data do not rule out the possibility of other variables being responsible for the change in behavior (Cooper, Heron, & Heward, 2007). Because there is no withdrawal of the intervention or replication of treatment effects, rival hypotheses based on factors that are not controlled by the teacher may have caused the changes in the behavior (e.g., the student may have matured over time, a new medication was administered, a parent has been working on the skill at home).

Marc's teachers decided to implement an intervention to increase his interactions during the 9:30 a.m. free play. Intervention consisted of seating Marc in proximity to two of his outgoing, friendly classmates, Sam and Mario, and praising the students for all interactions. Marc's interactions with his peers increased. However, Ms. Wharton wondered whether it was the new intervention (i.e., praise and environmental manipulation) that increased the peer interaction or if, perhaps, it might be the new action figures that Sam was bringing to school.

FIGURE 4–6
An AB Design for Evaluating the Effects of Intervention on Christine's Engagement During Drama Club (with and Without a Mean Line)



Given the lack of experimental control, the AB design is not considered acceptable for experimental research, but it is appropriate for monitoring student performance within teaching settings, particularly when proven teaching strategies are used. Sometimes, however, it is possible for teams to use a reversal design (ABAB), as described earlier when Jacob's teacher had the opportunity to assess his behavior under medication and without medication.

When teams are familiar with a student's typical learning patterns and are aware of various events that affect performance, they can usually judge treatment effects with a considerable degree of certainty. Because student performance is monitored before teaching occurs (during the baseline phase) and during different phases of teaching, the AB design provides an objective (although not scientifically conclusive) description of a student's behavior before, during, and after teaching.

If a student's performance during the baseline phase progresses in the same direction as is expected during the intervention, teachers have difficulty interpreting the intervention data. That is, unless the intervention has a great effect, it is difficult to judge whether the change in the data (from the baseline phase to the intervention phase) is the result of the intervention or simply a continuation of the trend seen during the baseline phase. Because of this difficulty, teachers typically wait until the baseline performance is relatively stable before starting intervention; however, in some instances, as discussed previously (e.g., dangerous behavior), teachers may not wait for a stable baseline or may not take a baseline measurement at all before beginning intervention.

An AB design is shown in Figure 4–6. This example shows an increase in Christine's engagement during drama club after intervention. Intervention consisted of having her sit closer to peers, having the teaching assistant move to the back of the room to do paperwork, teaching her peers how to communicate with her, and programming her communication device with vocabulary suited to the class.

Graphing Conventions

As described earlier, the main purpose of converting raw data to graphs is to provide a visual summary of the student's performance and progress. These interpretations guide program modification. The ability to accurately interpret graphs is enhanced by the use of certain graphing conventions. Four major graphing conventions are shown in Figure 4–4:

- *Broken vertical lines,* or phase changes, represent changes in the instructional program. These can include planned programmatic changes (e.g., a change from baseline to intervention measurement, a modification of the task analysis or materials, a change in the time or setting of the instruction, or a change in the reinforcement or prompt). Broken vertical lines can also be used to indicate situations or events that might indirectly or incidentally affect the student's performance. These events include, for example, changes in medication, staff changes, or a new student joining the class.
- *Broken horizontal lines* (criteria lines) can be used to indicate the criterion for the program, in other words, what level of performance is considered successful. This criterion should match the criterion stated in the behavioral objective of the program. Seeing this line on the graph gives the team a quick visual reminder of what the goal of the program is and where the student is in relation to the goal.
- *Connect data points* only for consecutive days within a phase. Data points should not be connected across phase change lines or across missed data days (e.g., a student absence, a missed session, holiday break). This allows a clear picture of the effect of the program change but also allows the team to note the gaps in instructional opportunities and its effect on performance.
- *Show the expected trend.* Teams can quickly note whether a graph represents a successful or an unsuccessful program effort by using an X to represent behavior that is to be decreased and dots to represent skill-building. This graphing convention also allows multiple data paths on one graph (e.g., increasing toy play and decreasing self-stimulatory behavior during free play).

Marc's teacher graphed his progress in increasing peer interactions during his 9:30 a.m. free play. Ms. Wharton decided that it would be helpful not just to note his progress for this social interaction skill, but also to analyze the impact of the intervention on Marc's behavior of grabbing toys from his peers. She drew a criterion line across the graph at 70% because the objective was to have Marc engage in appropriate peer interactions for 70% of the intervals during the free play (this percentage was based on her observations of other children in the class playing together). When Marc reached the criterion level for peer interactions and there was also a significant decrease in his grabbing of toys, Ms. Wharton decided that she should reduce the amount of praise she was giving to Marc and his peers so that it would be more typical of the frequency of praise she provided to other children in the class. To indicate this program change, she drew a broken vertical line on the graph. She was happy to note that the appropriate peer interactions remained high and that the grabbing behavior remained at zero (see Figure 4–4).

Visual Analysis

Visual inspection of a graph is the most frequently used strategy to evaluate the effects of classroom data (Alberto & Troutman, 2012; Hojnoski, Gischlar, & Missall, 2009). That is, certain characteristics and comparisons of the data paths (e.g., data from the baseline compared with data from the first phase of instruction) are examined to judge the effectiveness of the instructional strategy. Sometimes, the effects of a program are so strong that the impact is obvious by just looking at the graph. In these situations, when the graph itself lets you know all that you need to know, you are using visual analysis. For example, the top panel of Figure 4–6 shows a very distinct difference between the baseline and the intervention. It is easy to see on this graph a positive change both in the trend (an accelerating data path) and the level of change (performance of 10% and 20% during baseline, which then jumped to 40% and higher during intervention). As described above, however, an AB design is not an experimental design, and it is possible that confounding variables may be contributing to the change in behavior.

However, when such strong changes in behavior are not so obvious, teachers and researchers often find that their efforts in analyzing the trends on a graph are aided by several simple visual aids—mean line, aim line, and trend line.

Mean Lines

An easy visual addition to a graph to help analyze intervention effects is looking at the mean (i.e., the average) of each phase and comparing these lines. This can be done by calculating the mean of the data points in each phase and drawing a horizontal broken line that corresponds to that value on the ordinate scale (Alberto & Troutman, 2012). For example, we could calculate the average of the baseline for Figure 4–6 in the following manner:

1. Add the baseline data point values:

$$10\% + 20\% + 10\% = 40\%$$

2. Divide the total by the number of days:

$$\frac{40\%}{3} = 13.3\%$$

3. Draw a horizontal broken line across the baseline phase at 13%.

The same would be done for the intervention:

1. Add the intervention data point values:

$$40\% + 40\% + 40\% + 60\% + 60\% + 70\% + 70\% = 380\%$$

2. Divide the total by the number of days:

$$\frac{380\%}{7} = 54.28\%$$

3. Draw a horizontal broken line across the intervention phase at 54%.

The bottom panel of Figure 4–6 shows the difference in the two horizontal lines and is an additional visual representation of the success of the intervention. Mean lines can be particularly helpful in detecting change when the difference in the levels of behavior between phases is obvious.

Aim Lines

The aim line is a more sophisticated version of the criterion line that is calculated from a student's data collected during the baseline assessment or after several days of training. Usually, the aim line starts at the initial performance level of the student when you begin instruction and extends over the instructional period to the criterion level and date that you have set in the instructional objective (Farlow & Snell, 2005). This progress-monitoring visual aid is drawn onto a graph early in a program and allows the team to compare actual progress to their expectations for progress. Although they might look quite complicated, aim lines are simple to draw and easy to interpret. Aim lines result from connecting two points made by (a) the intersection of the mid-date and the mid-performance of the first three training days (or the last three baseline days) and (b) the intersection of the criterion performance with the goal date of accomplishment (i.e., the aim date).

In March, Jacob's team was concerned about his organizational skills and how they would affect him in fifth grade when following the schedule was even more important. Specifically, they wanted him to learn a four-step task analysis for following his schedule: (a) Look at his schedule and state the next activity, (b) gather the needed materials, (c) go to the appropriate location and participate, and (d) leave when done and return to his classroom. First, the team revised his schedule so that it reflected the 15 activities that he needed to do each day, including two bathroom breaks so that he would not skip going, which had resulted in accidents. Then, Ms. Bowers met with him and showed him the plan for the revised schedule to determine the changes that he might suggest. Jacob wanted the schedule to fit into his notebook and not be so obvious to his

classmates, and he did not want VELCRO® symbols because they were too "noisy," but he agreed that the new schedule would help him remember. The final version of his schedule consisted of 15 small picture/word symbols that were slipped into clear plastic pockets from left to right and top to bottom on one page. This way, the schedule symbols could be adjusted during his check-in each morning, if necessary. Then the team developed a teaching program and a data collection sheet and schedule (see Figure 4–7). Prior to recording the baseline data, Ms. Bowers gave Jacob the new schedule and various staff members (depending on where he was) recorded his performance over three days, withholding all prompts. The team wanted to determine whether the schedule alone would be enough. On his first day of the baseline period (see Figure 4–7), he used his schedule independently for 8 out of the 15 activities, which indicated that he needed instruction in addition to the schedule.

FIGURE 4–7
Jacob's Schedule Use Program and Data Collection Sheet

Student Name: Jacob

Days: Daily

Settings: Classroom, hallways, playground, lunchroom, gym, music room, computer lab, resource room, library

Instructional Cue: "Jacob, check your schedule."

Latency: 5 seconds (probes and teaching)

Baseline/Probe Method: Multiple opportunity task analytic assessment; don't repeat cue

Probe Schedule: Every Wednesday

Teaching Method: System of Least Prompts: verbal; verbal + gestural; verbal + physical

Recording Key:

Probe codes: ✓ independent on all steps, – incorrect/no response on one or more steps

Teach codes:
+ Attended to activity with prompting or independently (some of 4 steps needed prompting, some independent)
– Needed prompts on all steps

Materials: Visual schedule in his notebook, academic materials for scheduled activity in desk, classroom, or backpack

Objective: Jacob will use his schedule, identify the activity, get the needed materials, go to the activity and participate, and leave the activity when it is finished for all 15 daily activities

Date: 23 March baseline Staff: MM	Arrival, Check-in	A.M. Work	Language Arts	Guided*	Specialty**	Recess	Restroom	Math	Lunch	Science/Social Studies	Writers' Workshop	Shared Reading	Restroom	Check-out	Departure
1. Look at schedule, name activity.	✓	–	✓	–	✓	–	–	✓	–	✓	–	✓	✓	✓	✓
2. Get any needed materials.	✓	–	✓	–	✓	–	–	✓	–	✓	–	–	✓	✓	✓
3. Go to activity.	✓	–	✓	✓	✓	–	–	✓	–	✓	–	–	✓	✓	✓
4. Leave activity, return to class.	✓	–	✓	–	✓	–	–	✓	✓	✓	–	–	✓	✓	✓
Summary score: 8/15 = 53%	✓	–	✓	–	✓	–	–	✓	–	✓	–	–	✓	✓	✓

*Resource Room

**Monday: Music; Tuesday and Thursday: P.E.; Wednesday: Library; Friday: Computer

Note: Credit is given to Mandy McKee for this adaptation of her instructional program.

TABLE 4–5
Data for Jacob's Program for Using a Schedule

Date	Arrival, Check-In	Morning Work	Language Arts	Guided	Specialty	Recess	Restroom	Math	Lunch	Science/Social Studies		Writers' Workshop	Shared Reading	Restroom	Checkout	Departure	Total Independent	Percentage Correct During the Day
23-Mar	✓	–	✓	–	✓	–	–	✓	–	✓		–	–	✓	✓	✓	8	53
24-Mar	✓	–	✓	–	✓	–	✓	✓	–	✓		✓	–	✓	–	✓	9	60
25-Mar	✓	–	–	–	✓	–	–	✓	–	✓		✓	–	✓	–	✓	7	47
26-Mar	+	–	–	–	+	–	+	+	–	+	Early release						5	50
27-Mar	+	+	+	–	+	–	–	+	–	+		+	–	+	–	+	9	60
30-Mar	+	–	+	–	+	+	+	+	–	+		+	–	+	+	+	11	73
31-Mar	+	–	–	+	+	–	+	+	+	+		+	–	+	+	+	11	73
1-Apr	✓	–	✓	✓	✓	✓	–	✓	–	✓		✓	–	✓	✓	✓	11	73
3-Apr	+	Late arrival				–	+	+	–	+		+	–	+	+	+	8	73
Spring break: 4-Apr through 13-Apr																		
14-Apr	+	–	+	+	+	+	+	+	+	+		+	–	+	–	+	12	80
15-Apr	✓	–	–	–	✓	–	✓	✓	✓	✓		✓	–	✓	–	✓	9	60
16-Apr	+	–	–	+	+	+	+	+	+	+		+	–	+	–	+	11	73
17-Apr	+	Late arrival								+	+	+	–	+	–	+	6	75
20-Apr	+	Late arrival						+	+	–	+	–	–	+	+	+	7	70
21-Apr	+	Field trip +									+	+	–	+	+	+	7	88
22-Apr	✓	–	+	–	✓	–	✓	✓	✓	✓		✓	–	✓	✓	✓	11	73
23-Apr	+	–	–	+	+	+	+	+	+	+		+	–	+	–	+	11	73
24-Apr	+	–	–	+	+	+	+	+	+	+		+	–	+	–	+	11	73

Shaded data are probe data.

Probe codes: ✓ Independent on all steps, - Incorrect/No response on one or more of the four task steps

Teach codes: + Attended to activity with prompting or independently (some of four steps needed prompting, some were independent)
– Needed prompts for all steps

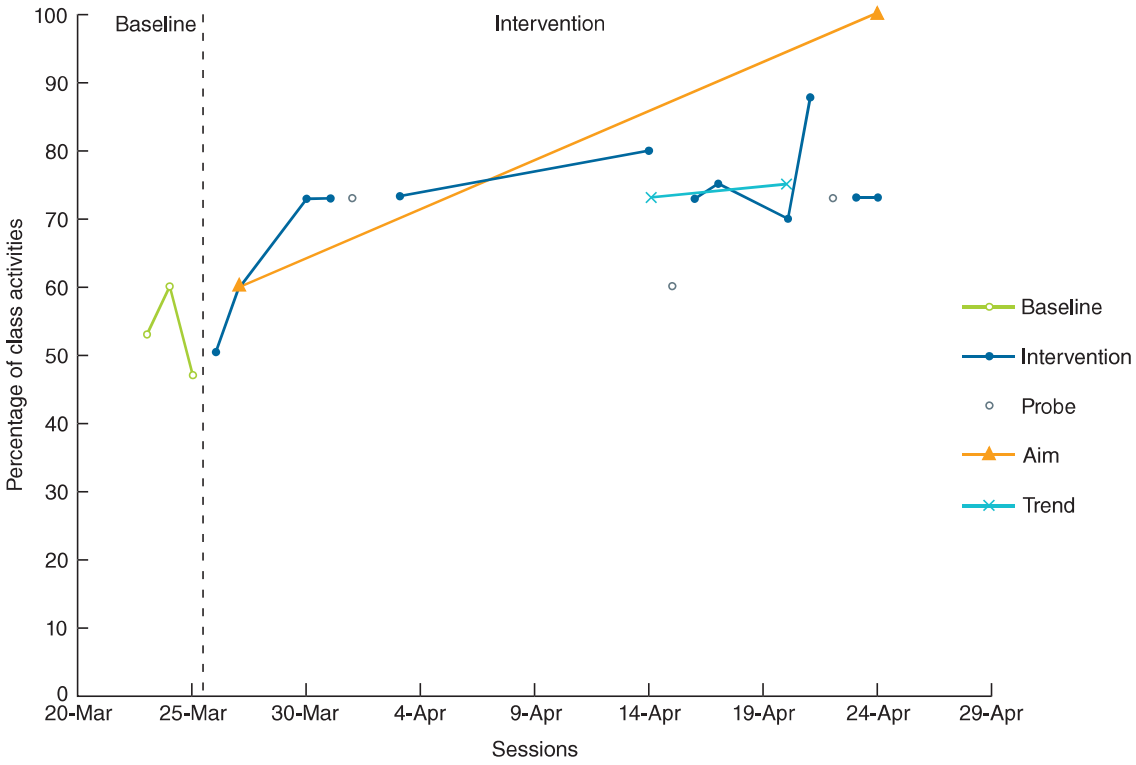
Note: Credit is given to Mandy McKee for this adaptation of her instructional program.

Table 4–5 shows Jacob's data for the entire program. Gray lines are probes, while unshaded lines are training data. Figure 4–8 shows a graph of these data. Jacob's teachers drew an aim line on the graph at the beginning of his program for the purpose of guiding their judgments about his progress. Look at the aim line in the graph and let's review how his teachers drew it on the graph.

Jacob's first three teaching data points in Table 4–5 show that he scored 50%, 60%, and 73%, respectively, during the first 3 days of teaching (March 26, 27, and 30); this was a gradual improvement over his baseline performance.

Ms. Fuentes, Jacob's special education teacher, was in charge of entering the data onto a graph for the team. After the first three training data points had been graphed, she drew the aim line. She began by setting the anchor (beginning point) for the aim line at the intersection of the second day of teaching, or mid-date (March 27), and his mid-performance, which was 60% (i.e., the middle value of the three percentages). The aim line endpoint was set at the intersection of the criterion performance that was expected

FIGURE 4–8
Percentage of Class Activities Attended by Jacob During the School Day



Note: Credit is given to Mandy McKee for this adaptation of her instructional program.

for Jacob (i.e., 100%) and the goal date for learning pennies (April 24). Jacob's teacher then drew a straight line connecting these two points.

The selection of the aim criterion reflects the team's expectations for the student. Jacob's team wanted him to master schedule use and his baseline performance indicated that he was independent about half of the time. Thus, they set his criterion at 100%. The selection of an aim date is influenced by several factors, including the school's typical evaluation periods (e.g., every nine weeks), the urgency (or timeline) for learning the skill, the difference between current performance and the criterion, and the speed of learning on similar tasks in the recent past. Aim dates should not be excessively distant from implementation dates and may be set to correspond to the more frequent marking periods used in general education (e.g., 9 to 12 weeks). Jacob's team chose the third week in April, as they hoped to have this skill in place before the fourth graders started their state assessment testing.

Trend Lines

A trend line roughly averages the direction and slope of a student's performance when the performance is uneven, variable, or difficult to interpret by just looking at the graph. A trend indicates the direction of graphed data, as well as the slope, or steepness, of the data path. A trend can be of three general types:

1. *Ascending trends* have an upward slope on a graph and indicate improvement (or learning) when the behavior graphed is a skill or adaptive behavior. (If the goal of the program is to reduce a behavior, the ascending trend would be interpreted as regression, or deterioration.)
2. *Flat trends* have either no slope or a very slight upward or downward slope. When the behavior graphed is a skill or an adaptive behavior, flat trends indicate no learning or improvement.

3. *Descending trends* have a downward slope and indicate regression, or deterioration, when the behavior graphed is a skill or an adaptive behavior. (If the goal of the program is to reduce a behavior, the downward slope would be interpreted as indicating improvement.)

When the trend of graphed teaching data is not obvious or uniform, trend lines help teachers summarize and interpret the fluctuating, or variable, nature of performance (Tawney & Gast, 1984). Some researchers have defined an ascending slope as 30° or more in the positive direction and a descending slope as 30° or more in the negative direction (DeProspero & Cohen, 1979).

When the trend is obvious, there is no need to draw a trend line, but when a student's progress is below the aim line for three out of five consecutive data points and the trend is not obvious, a teacher should pencil in a trend line to define the trend. If the line indicates that the trend on a target skill is flat or descending (for skill-building programs), then the team should further analyze the data and other relevant information to decide whether specific program modifications are needed.

Jacob stayed above the aim line until spring break (April 4). When he returned, his performance varied, but was below the aim line even after a week of teaching. Ms. Fuentes drew a trend line to help summarize his performance (see Figure 4–8).

There are several ways to draw trend lines, including the "quickie split-middle trend line" (White & Haring, 1980), which can be drawn simply and clarifies the general direction of change in the data, as well as the relative rate of change (reflected in the slope). Although teachers may collect test (probe) data intermittently and add them to the same graph using differently coded lines or points, teaching or training data constitute the primary information used to make judgments about day-to-day progress (Browder, 1991; Farlow & Snell, 1994; Haring, Liberty, & White, 1980). Teachers should draw a trend line using the following six steps (see Figure 4–8):

1. Take the last 6 to 10 days of teaching data collected and draw a vertical line to divide the data in half.
2. Look at the first half of the data (the first three, four, or five data points) and locate the middle date. Draw a small vertical line through this data point.
   - If an odd set of data points is considered (e.g., three or five), draw the line through the middle date.
   - If an even set of data points is considered (e.g., four), just sketch a pencil line between the dates for the second and third data points.
3. Look at the middle performance level and draw a short horizontal line through this data point. The middle performance level is not the average of the performances for the data points in the set of three to five data points. It is simply the middle performance value.
   - For an odd number of data points, select the middle value (e.g., for 10%, 15%, and 12%, the middle performance value is 12%).
   - For an even number of data points, select a value halfway between the two middle data points (e.g., for 15%, 10%, 15%, and 11%, the middle value between 11 and 15 is 13%).
4. Extend the two lines until they intersect.
5. Repeat this process for the other half of the data.
6. Then draw a line connecting the two points of intersection from both halves of the data.

Once the trend line is drawn, visually judge whether the trend is ascending, flat, or descending. If the trend is flat or descending, teachers should hypothesize or determine why and make program modifications on the basis of the hypotheses or explanations. If the trend is ascending but not fast enough, teachers may (a) make changes in the program (e.g., modify the materials or prompts, change the reinforcer) to try to

speed progress, or (b) adjust the aim line by lowering the criterion or moving the date further into the future. Alternately, teachers may do nothing to the program and look for other explanations for the student's reduced progress.

Jacob's trend line was drawn using six days of training data starting with April 3 and ending with April 20. Ms. Fuentes divided the data into two groups: the first three data points and the last three. For the first 3 days (April 3, 14, and 15: 73%, 80%, 60%, respectively), the middle date was April 14 and the middle value was 73%. For the last 3 days (April 16, 17, and 20: 73%, 75%, 70%, respectively), the middle date was April 17 and the middle value was 73%. She connected the two points that resulted by drawing a line. The trend was very slightly ascending, but practically flat. The team members discussed their options because Jacob had stopped making progress. They could lower the criterion from 100%, move the aim date to May, or improve the program. They decided to keep the criterion at 100% because this was a skill that he really needed. They also decided to move the aim date to the end of May, giving more time and a more gradually ascending aim line. Finally, they decided to improve the program in two ways: (a) As a way to help motivate Jacob, they added two other classmates to the program who also had problems following the schedule, and (b) they added high-fives to the reinforcement from each teacher following steps 1 through 3 and following Step 4 if completed correctly.

Teams may decide to use computer-generated graphs instead of hand-drawn ones. Graphs can be fairly easily drawn and data added using Microsoft Excel and the step-by-step guide of Carr and Burkholder (1998). In addition, trend lines and aim lines can be constructed, although a somewhat different version of the quickie split-middle trend line (White & Haring, 1980) is produced with this program.

LEARNING OUTCOME SUMMARIES

5.01 Why Measure Student Behavior?

Learning Outcome

1. Describe what is meant by an evidence base.

The term *evidence-based practice* refers to educational programs or instructional procedures that have been experimentally determined to reliably produce successful student outcomes. A strategy that is considered to have an evidence base is one that has sufficient numbers of high-quality research studies published in peer-reviewed journals.

2. Describe the difference between outcome and process measures and why each helps educators to be accountable.

Process measures focus on small units of behavior such as the student's performance on steps of a task analysis. These measures are important when we evaluate a student's progress in an instructional program. If the data indicate that the student is not progressing, educators will need to make data-based decisions regarding program modifications. *Outcome measures* provide a broader view of student performance. These measures offer information regarding the general effects of a program on a person's quality of life. Examples of outcomes measure include friendships, membership in the classroom, time spent with peers in social situations, and increased self-determination. These measures are important because they give us the opportunity to evaluate the impact of the educational program on the student's, and his or her family's, quality of life.

5.02 Foundations of Meaningful Measurement

Learning Outcome

1. Identify the five criteria for evaluating the validity and importance of behavior changes.

The five criteria that are used to evaluate the validity and importance of behavior changes in a student are statistical significance (Was the difference in behavior beyond what might be

expected by chance?), clinical significance (Will the change in behavior allow the student to better function in society?), social validity (Are the changes in behavior important to the individual and make a difference in his or her life?), internal validity (Are we confident that the changes are due to the intervention?), quality of life (Will the change in the behavior improve the student's vision of a quality life?), and treatment integrity (Are the intervention procedures implemented in a way that is specified in the treatment plan?).

2. Discuss why measurement should be contextually appropriate as well as accurate and reliable.

Measurement systems must be developed that allow accurate measurement of the behavior, but also that fit into the environment in which it is being used. Professionals need to choose data collection strategies that are unobtrusive but that still provide sufficient information to determine if an intervention is effective. Measurement must also be accurate and reliable; this is accomplished by creating an *operational definition of the target behavior,* that is, a description of the behavior that is observable and measurable. Without defining the behavior we cannot be sure if everyone is observing and addressing the same behavior, and cannot determine if a program is successful or needs modifications.

5.03 Quantitative Measures

Learning Outcome

Identify and describe seven different measures of student performance.

Frequently used measures of student performance include

1. Permanent product: direct measurement of the lasting and concrete results of a target behavior
2. Frequency recording: the number of times a behavior occurs within a specified period
3. Percentage: the correct number of behavior compared with the number of opportunities (or intervals)
4. Rate: the frequency of a behavior and its relationship to time, expressed as a ratio
5. Duration: the total amount of time in which a targeted behavior occurs in a specified observation
6. Task analytic measurement: a record of the performance of each step in a sequence of behaviors that make up a task
7. Interval recording: a record of the occurrence of a behavior within each of the time intervals within a single observation. Three interval strategies are whole interval, partial interval, and momentary time sampling.

5.04 Organizing Student Performance

Learning Outcome

1. Describe the various elements of a graph.

A graph is made up of two axes: the abscissa (*x*-axis) and the ordinate (*y*-axis). The abscissa is the horizontal line that usually represents the time frame of a measurement. The ordinate is the vertical line and is labeled with the target behavior being measured. Once data are converted into a single numeral, this is plotted on the graph.

2. Discuss factors that influence how frequently student data should be collected.

Just how often an educator should take data remains a subject of debate. Although there are no definitive answers to this question, several guidelines can assist in deciding how frequently to collect data:

1. Higher priority objectives warrant daily data collection.
2. Lower priority objectives may be evaluated less frequently.
3. Implementation of a new program should initially be measured every teaching session until progress is consistent.
4. Data that show progress as planned may be evaluated less frequently.
5. Data that do not show progress as planned or that are variable should be measured daily.

5.05 Data Analysis for Better Decision-Making

Learning Outcome

1. Discuss the difference between probe (testing) and teaching data, and the uses for each.

Probing (or testing) means that a person's performance is checked under criterion conditions (i.e., conditions that as closely as possible use natural contexts, cues, and consequences). When conducting a probe, the educator typically provides no prompting or teaching assistance, no reinforcement for task success or improvement, and no corrections. The goal of testing is to learn about a student's current performance under criterion conditions, and not to teach the student. In contrast, *teaching data* can be recorded whenever the student is taught. Under teaching conditions data are recorded while a student is participating in the instructional intervention. Learning is the goal of teaching, so conditions are planned to promote improvement in performance and to advance the student through the various stages of learning.

2. Describe how a baseline–intervention comparison can be useful to an educator, and the limitations of this type of analysis.

The baseline–intervention (or AB design) is referred to as a non-experimental design because no conclusive demonstration of a cause–effect, or functional, relationship between the intervention and the observed changes in behavior is possible. However, this comparison is appropriate for monitoring student performance within the educational settings, particularly when proven teaching strategies are used.

3. Describe the various graphing conventions that allow for a visual summary of performance data.

There are several graphing conventions that support the visual analysis of a graph. These include (a) broken vertical lines that represent changes in the instructional program, (b) broken horizontal lines (a criterion line) to indicate the criterion for the program, and (c) connecting data points only for consecutive days within a phase.

SUGGESTED ACTIVITIES

1. Select one student in your class and examine the current measurement strategies used for each IEP objective.
   a. For each objective that includes a measurement strategy, consider whether that strategy is the most appropriate one to use. Would a different strategy be more meaningful?
   b. For each objective that does not include a measurement strategy, describe an appropriate and manageable strategy that could be used.
2. For this same student, design a graph for each instructional objective not currently displayed in graphic format.