CHAPTER 5

Improving and Assessing the Quality of Behavioral Measurement

Key Terms

accuracy
believability
calibration
continuous measurement
direct measurement
discontinuous measurement
exact count-per-interval IOA
indirect measurement
interobserver agreement (IOA)
interval-by-interval IOA
mean count-per-interval IOA
mean duration-per-occurrence IOA
measurement bias
naive observer
observed value
observer drift
observer reactivity
reliability
scored-interval IOA
total count IOA
total duration IOA
trial-by-trial IOA
true value
unscored-interval IOA
validity

Behavior Analyst Certification Board® BCBA® & BCABA® Behavior Analyst Task List, © Third Edition

	Content Area 6: Measurement of Behavior
6-4	Select the appropriate measurement procedure given the dimensions of the behavior and the logistics of observing and recording.
6-5	Select a schedule of observation and recording periods.
6-14	Use various methods of evaluating the outcomes of measurement procedures, such as interobserver agreement, accuracy, and reliability.

The data obtained by measuring behavior are the primary material with which behavioral researchers and practitioners guide and evaluate their work. Applied behavior analysts measure socially significant behaviors to help determine which behaviors need to be changed, to detect and compare the effects of various interventions on behaviors targeted for change, and to evaluate the acquisition, maintenance, and generalization of behavior changes.

Because so much of what the behavior analyst does either as a researcher or practitioner depends on measurement, concerns about the legitimacy of the data it produces must be paramount. Do the data meaningfully reflect the original reason(s) for measuring the behavior? Do the data represent the true extent of the behavior as it actually occurred? Do the data provide a consistent picture of the behavior? In other words, can the data be trusted?

Chapter 4 identified the measurable dimensions of behavior and described the measurement methods most often used in applied behavior analysis. This chapter focuses on improving and assessing the quality of behavioral measurement. We begin by defining the essential indicators of trustworthy measurement: validity, accuracy, and reliability. Next, common threats to measurement are identified and suggestions for combating these threats are presented. The chapter's final sections detail procedures for assessing the accuracy, reliability, and believability of behavioral measurement.

Indicators of Trustworthy Measurement

> Three friends—John, Tim, and Bill—took a bicycle ride together. At the end of the ride John looked at his handlebar-mounted bike computer and said, "We rode 68 miles. Excellent!" "My computer shows 67.5 miles. Good ride, fellas!" Tim replied. As he dismounted and rubbed his backside, the third biker, Bill, said, "Gee whiz, I'm sore! We must've ridden 100 miles!" A few days later, the three friends completed the same route. After the second ride, John's computer showed 68 miles, Tim's computer read 70 miles, and Bill, because he wasn't quite as sore as he was after the first ride, said they had ridden 90 miles. Following a third ride on the same country roads, John, Tim, and Bill reported distances of 68, 65, and 80 miles, respectively.

How trustworthy were the measures reported by the three bicyclists? Which of the three friends' data would be most usable for a scientific account of the miles they had ridden? To be most useful for science, measurement must be valid, accurate, and reliable. Were the three friends' measurements characterized by validity, accuracy, and reliability?

Validity

Measurement has **validity** when it yields data that are directly relevant to the phenomenon measured and to the reason(s) for measuring it. Determining the validity of measurement revolves around this basic question: Was a relevant dimension of the behavior that is the focus of the investigation measured directly and legitimately?

Did the measurements of miles ridden by the three bicyclists have validity? Because the bikers wanted to know how far they had ridden each time, the number of miles ridden was a relevant, or valid, dimension of their riding behavior. Had the bikers' primary interest been how long or how fast they had ridden, the number of miles ridden would not have been a valid measure. John and Tim's use of their bike computers to measure directly the miles they rode was a valid measure. Because Bill used an indirect measure (the relative tenderness of his backside) to determine the number of miles he had ridden, the validity of Bill's mileage data is suspect. A direct measure of the actual behavior of interest will always possess more validity than an indirect measure, because a direct measure does not require an inference about its relation to the behavior of interest, whereas an indirect measure always requires such an inference. Although soreness may be related to the distance ridden, because it is also influenced by such factors as the time on the bike saddle, the roughness of the road, riding speed, and how much (or little) the person has ridden recently, soreness as a measure of mileage has little validity.

Valid measurement in applied behavior analysis requires three equally important elements: (a) measuring directly a socially significant target behavior (see Chapter 3), (b) measuring a dimension (e.g., rate, duration) of the target behavior relevant to the question or concern about the behavior (see Chapter 4), and (c) ensuring that the data are representative of the behavior's occurrence under conditions and during times that are most relevant to the question or concern about the behavior. When any of these elements are suspect or lacking—no matter how technically proficient (i.e., accurate and reliable) was the measurement that produced the data—the validity of the resultant data are compromised, perhaps to the point of being meaninglessness.

Accuracy

When used in the context of measurement, **accuracy** refers to the extent to which the **observed value,** the quantitative label produced by measuring an event, matches the true state, or true value, of the event as it exists in nature (Johnston & Pennypacker, 1993a). In other words, measurement is accurate to the degree that it corresponds to the true value of the thing measured. A **true**

value is a measure obtained by procedures that are independent of and different from the procedures that produced the data being evaluated and for which the researcher has taken "special or extraordinary precautions to insure that all possible sources of error have been avoided or removed" (p. 136).

How accurate were the three bikers' measures of miles ridden? Because each biker obtained a different measure of the same event, all of their data could not be accurate. Skeptical of the training miles the three cyclists were claiming, a friend of theirs, Lee, drove the same country roads with a Department of Transportation odometer attached to the back bumper of his car. At the end of the route the odometer read 58 miles. Using the measure obtained by the DOT odometer as the true value of the route's distance, Lee determined that none of the three cyclists' measures were accurate. Each rider had overestimated the true mileage.

By comparing the mileage reported by John, Tim, and Bill with the true value of the route's distance, Lee discovered not only that the riders' data were inaccurate, but also that the data reported by all three riders were contaminated by a particular type of measurement error called measurement bias. **Measurement bias** refers to nonrandom measurement error; that is, error in measurement that is likely to be in one direction. When measurement error is random, it is just as likely to overestimate the true value of an event as it is to underestimate it. Because John, Tim, and Bill consistently overestimated the actual miles they had ridden, their data contained measurement bias.

Reliability

Reliability describes the extent to which a "measurement procedure yields the same value when brought into repeated contact with the same state of nature" (Johnston & Pennypacker, 1993a, p. 138). In other words, reliable measurement is consistent measurement. Like validity and accuracy, reliability is a relative concept; it is a matter of degree. The closer the values obtained by repeated measurement of the same event are to one another, the greater the reliability. Conversely, the more observed values from repeated measurement of the same event differ from one another, the less the reliability.

How reliable were the bicyclists' measurements? Because John obtained the same value, 68 miles, each time he measured the same route, his measurement had complete reliability. Tim's three measures of the same ride—67.5, 70, and 65 miles—differed from one another by as much as 5 miles. Therefore, Tim's measurement was less reliable than John's. Bill's measurement system was the least reliable of all, yielding values for the same route ranging from 80 to 100 miles.

Relative Importance of Validity, Accuracy, and Reliability

Behavioral measurement should provide legitimate data for evaluating behavior change and guiding research and treatment decisions. Data of the highest quality (i.e., data that are most useful and trustworthy for advancing scientific knowledge or for guiding data-based practice) are produced by measurement that is valid, accurate, and reliable (see Figure 5.1). Validity, accuracy, and reliability are relative concepts; each can range from high to low.

Measurement must be both valid and accurate for the data to be trustworthy. If measurement is not valid, accuracy is moot. Accurately measuring a behavior that is not the focus of the investigation, accurately measuring an irrelevant dimension of the target behavior, or accurately measuring the behavior under circumstances or at times not representative of the conditions and times relevant to the analysis will yield invalid data. Conversely, the data obtained from measuring a meaningful dimension of the right behavior under the relevant circumstances and times is of little use if the observed values provide an inaccurate picture of the behavior. Inaccurate measurement renders invalid the data obtained by otherwise valid measurement.

Reliability should never be confused with accuracy. Although John's bicycle computer provided totally reliable measures, it was also totally inaccurate.

> Concern about the reliability of data in the absence of a prior interest in their accuracy suggests that reliability is being mistaken for accuracy. The questions for a researcher or someone who is reading a published study is not, "Are the data reliable?" but "Are the data accurate?" (Johnston & Pennypacker, 1993a, p. 146)

If accuracy trumps reliability—and it does—why should researchers and practitioners be concerned with the reliability of measurement? Although high reliability does not mean high accuracy, poor reliability reveals problems with accuracy. Because Tim and Bill's measurements were not reliable, we know that at least some of the data they reported could not be accurate, knowledge that could and should lead to checking the accuracy of their measurement tools and procedures.

Highly reliable measurement means that whatever degree of accuracy (or inaccuracy) exists in the measurement system will be revealed consistently in the data. If it can be determined that John's computer reliably obtains observed values higher than the true values by a constant amount or proportion, the data could be adjusted to accommodate for that constant degree of inaccuracy.

The next two sections of the chapter describe methods for combating common threats to the validity, accuracy, and reliability of behavioral measurement.

Figure 5.1 Measurement that is valid, accurate, and reliable yields the most trustworthy and useful data for science and science-based practice.

Measurement that is . .

Valid	Accurate	Reliable	. . . yields data that are . . .
Yes	Yes	Yes	. . . most useful for advancing scientific knowledge and guiding data-based practice.
No	Yes	Yes	. . . meaningless for the purposes for which measurement was conducted.
Yes	No	Yes	. . . always wrong.[1]
Yes	Yes	No[2]	. . . sometimes wrong.[3]

1. If adjusted for consistent measurement error of standard size and direction, inaccurate data may still be usable.
2. If the accuracy of every datum in a data set can be confirmed, reliability is a moot point. In practice, however, that is seldom possible; therefore, knowing the consistency with which a valid and accurate measurement system has been applied contributes to the level of confidence in the overall trustworthiness of the data set.
3. User is unable to separate the good data from the bad.

Threats to Measurement Validity

The validity of behavioral data is threatened when measurement is indirect, when the wrong dimension of the target behavior is measured, or when measurement is conducted in such a way that the data it produces are an artifact of the actual events.

Indirect Measurement

Direct measurement occurs when "the phenomenon that is the focus of the experiment is exactly the same as the phenomenon being measured" (Johnston & Pennypacker, 1993a, p. 113). Conversely, **indirect measurement** occurs when "what is actually measured is in some way different from" the target behavior of interest (Johnston & Pennypacker, 1993a, p. 113). Direct measurement of behavior yields more valid data than will indirect measurement. This is because indirect measurement provides secondhand or "filtered" information (Komaki, 1998) that requires the researcher or practitioner to make inferences about the relationship between the event that was measured and the actual behavior of interest.

Indirect measurement occurs when the researcher or practitioner measures a proxy, or stand-in, for the actual behavior of interest. An example of indirect measurement would be using children's responses to a questionnaire as a measure of how often and well they get along with their classmates. It would be better to use a direct measure of the number of positive and negative interactions among the children. Using a student's score on a standardized math achievement test as an indicator of her mastery of the math skills included in the school's curriculum is another example of indirect measurement. Accepting the student's score on the achievement test as a valid reflection of her ability with the school's curriculum would require an inference. By contrast, a student's score on a properly constructed test consisting of math problems from recently covered curriculum content is a direct measure requiring no inferences about what it means with respect to her performance in the curriculum.

Indirect measurement is usually not an issue in applied behavior analysis because meeting the applied dimension of ABA includes the targeting and meaningful (i.e., valid) measurement of socially significant behaviors. Sometimes, however, the researcher or practitioner has no direct and reliable access to the behavior of interest and so must use some form of indirect measurement. For example, because researchers studying adherence to medical regimens cannot directly observe and measure patients' behavior in their homes, they rely on self-reports for their data (e.g., La Greca & Schuman, 1995).[1]

Indirect measurement is sometimes used to make inferences about private events or affective states. For example, Green and Reid (1996) used direct measures of smiling to represent "happiness" by persons with profound multiple disabilities. However, research on private events does not necessarily involve indirect measurement. A research participant who has been trained to observe his own private events is measuring the behavior of interest directly (e.g., Kostewicz, Kubina, & Cooper, 2000; Kubina, Haertel, & Cooper, 1994).

[1]Strategies for increasing the accuracy of self-reports can be found in Critchfield, Tucker, and Vuchinich (1998) and Finney, Putnam, and Boyd (1998).

Whenever indirect measurement is used, it is the responsibility of the researcher to provide evidence that the event measured directly reflects, in some reliable and meaningful way, something about the behavior for which the researcher wishes to draw conclusions (Johnston & Pennypacker, 1993a). In other words, it is incumbent upon the researcher to provide a convincing case for the validity of her data. Although it is sometimes attempted, the case for validity cannot be achieved by simply attaching the name of the thing one claims to be measuring to the thing actually measured. With respect to that point, Marr (2003) recounted this anecdote about Abraham Lincoln:

> "Sir, how many legs does this donkey have?"
> "Four, Mr. Lincoln."
> "And how many tails does it have?"
> "One, Mr. Lincoln."
> "Now, sir, what if we were to call a tail a leg; how many legs would the donkey have?"
> "Five, Mr. Lincoln."
> "No sir, for you cannot make a tail into a leg by calling it one." (pp. 66–67)

Measuring the Wrong Dimension of the Target Behavior

The validity of behavioral measurement is threatened much more often by measuring the wrong dimension of the behavior of interest than it is by indirect measurement. Valid measurement yields data that are relevant to the questions about the behavior one seeks to answer through measurement. Validity is compromised when measurement produces values for a dimension of the behavior ill suited for, or irrelevant to, the reason for measuring the behavior.

Johnston and Pennypacker (1980) provided an excellent example of the importance of measuring a dimension that fits the reasons for measurement. "Sticking a ruler in a pot of water as the temperature is raised will yield highly reliable measures of the depth of the water but will tell us very little about the changing temperature" (p. 192). While the units of measurement on a ruler are well suited for measuring length, or in this case, depth, they are not at all valid for measuring temperature. If the purpose of measuring the water is to determine whether it has reached the ideal temperature for making a pot of tea, a thermometer is the correct measurement tool.

If you are interested in measuring a student's academic endurance with oral reading, counting the number of correct and incorrect words read per minute without measuring and reporting the total time that the student read will not provide valid data on endurance. Number of words read per minute alone does not fit the reason for measuring reading (i.e., academic endurance). To measure endurance, the practitioner would need to report the duration of the reading period (e.g., 30 minutes). Similarly, measuring the percentage of trials on which a student makes a correct response will not provide valid data for answering questions about the student's developing fluency with a skill, whereas measuring the number of correct responses per minute and the changing rates of responding (celeration) would.

Measurement Artifacts

Directly measuring a relevant dimension of a socially significant target behavior does not guarantee valid measurement. Validity is reduced when the data—no matter how accurate or reliable they are—do not give a meaningful (i.e., valid) representation of the behavior. When data give an unwarranted or misleading picture of the behavior because of the way measurement was conducted, the data are called an *artifact.* As introduced in Chapter 4, a *measurement artifact* is something that appears to exist because of the way it is measured. Discontinuous measurement, poorly scheduled measurement periods, and using insensitive or limiting measurement scales are common causes of measurement artifacts.

Discontinuous Measurement

Because behavior is a dynamic and continuous phenomenon that occurs and changes over time, continuous measurement is the gold standard in behavioral research. **Continuous measurement** is measurement conducted in a manner such that all instances of the response class(es) of interest are detected during the observation period (Johnston & Pennypacker, 1993a). **Discontinuous measurement** describes any form of measurement in which some instances of the response class(es) of interest may not be detected. Discontinuous measurement—no matter how accurate and reliable—may yield data that are an artifact.

A study by Thomson, Holmber, and Baer (1974) provides a good demonstration of the extent of artifactual variability in a data set that may be caused by discontinuous measurement. A single, highly experienced observer used three different procedures for scheduling time sampling observations to measure the behavior of four subjects (two teachers and two children) in a preschool setting during 64-minute sessions. Thomson and colleagues called the three time sampling procedures contiguous, alternating, and sequential. With each time sampling procedure, one-fourth of the observer's time (i.e., 16 minutes) was assigned to each of the four subjects.

When the contiguous observation scheduled was used, the observer recorded the behavior of Subject 1 throughout the first 16 minutes of the session, recorded the behavior Subject 2 during the second 16 minutes, and so on until all four students had been observed. In the alternating mode, Subjects 1 and 2 were observed in alternating intervals during the first half of the session, and Subjects 3 and 4 were observed in the same fashion during the last half of the session. Specifically, Student 1 was observed during the first 4 minutes, Subject 2 during the next 4 minutes, Subject 1 during the next 4 minutes, and so on until 32 minutes had expired. The same procedure was then used for Students 3 and 4 during the last 32 minutes of the session. The sequential approach systematically rotated the four subjects through 4-minute observations. Subject 1 was observed during the first 4 minutes, Subject 2 during the second 4 minutes, Subject 3 during the third 4 minutes, and Subject 4 during the fourth 4 minutes. This sequence was repeated four times to give the total of 64 minutes of observation.

To arrive at the percentage of artifactual variance in the data associated with each time sampling schedule, Thomson and colleagues (1974) compared the observer's data with "actual rates" for each subject produced by continuous measurement of each subject for the same 64-minute sessions. Results of the study showed clearly that the contiguous and alternating schedules produced the most unrepresentative (and therefore, less valid) measures of the target behaviors (often more than 50% variance from continuous measurement), whereas sequential sampling procedure produced results that more closely resembled the data obtained through continuous recording (from 4 to 11% variance from continuous measurement).

In spite of its inherent limitations, discontinuous measurement is used in many studies in applied behavior analysis in which individual observers measure the behavior of multiple subjects within the same session. Minimizing the threat to validity posed by discontinuous measurement requires careful consideration of when observation and measurement periods should be scheduled. Infrequent measurement, no matter how accurate and reliable it is, often yields results that are an artifact. Although a single measure reveals the presence or absence of the target behavior at a given point in time, it may not be representative of the typical value for the behavior.[2] As a general rule, observations should be scheduled on a daily or frequent basis, even if for only brief periods.

Ideally, all occurrences of the behavior of interest should be recorded. However, when available resources preclude continuous measurement throughout an observation period, the use of sampling procedures is necessary. A sampling procedure may be sufficient for decision making and analysis if the samples represent a valid approximation of the true parameters of the behavior of interest. When measurement cannot be continuous throughout an observation period, it is generally preferable to sample the occurrence of the target behavior for numerous brief observation intervals that are evenly distributed throughout the session than it is to use longer, less frequent intervals (Thomson et al., 1974; Thompson, Symons, & Felce, 2000). For example, measuring a subject's behavior in thirty 10-second intervals equally distributed within a 30-minute session will likely yield more representative data than will observing the person for a single 5-minute period during the half hour.

Measuring behavior with observation intervals that are too short or too long may result in data that grossly over- or underestimate the true occurrence of behavior. For example, measuring off-task behavior by partial-interval recording with 10-minute intervals may produce data that make even the most diligent of students appear to be highly off task.

Poorly Scheduled Measurement Periods

The observation schedule should be standardized to provide an equal opportunity for the occurrence or nonoccurrence of the behavior across sessions and consistent environmental conditions from one observation session to the next. When neither of these requirements is met, the resultant data may not be representative and may be invalid. If observation periods are scheduled at times when and/or places where the frequency of behavior is atypical, the data may not represent periods of high or low responding. For example, measuring students' being on-task during only the first 5 minutes of each day's 20-minute cooperative learning group activity may yield data that make on-task behavior appear higher than it actually is over the entire activity.

When data will be used to assess the effects of an intervention or treatment, the most conservative observation times should be selected. That is, the target behavior should be measured during those times when their frequency of occurrence is most likely to be different from the desired or predicted outcomes of the treatment. Measurement of behaviors targeted for reduction should occur during times when those behaviors are most likely to occur at their highest response rates. Conversely, behaviors targeted for increase should be measured when high-frequency responding is least likely. If an intervention is not planned—as might be the case in a descriptive study—it is important to select the observation times most likely to yield data that are generally representative of the behavior.

[2]Single measures, such as pretests and posttests, can provide valuable information on a person's knowledge and skills before and after instruction or treatment. The use of *probes*, occasional but systematic measures, to assess maintenance and generalization of behavior change is discussed in Chapter 28.

Insensitive and/or Limited Measurement Scales

Data that are artifacts may result from using measurement scales that cannot detect the full range of relevant values or that are insensitive to meaningful changes in behavior. Data obtained with a measurement scale that does not detect the full range of relevant performances may incorrectly imply that behavior cannot occur at levels below or above obtained measures because the scale has imposed an artificial floor or ceiling on performance. For example, measuring a student's oral reading fluency by giving him a 100-word passage to read in 1 minute may yield data that suggest that his maximum performance is 100 wpm.

A measurement scale that is over- or undersensitive to relevant changes in behavior may produce data that show misleadingly that meaningful behavior change has (or has not) occurred. For example, using a percentage measure scaled in 10% increments to evaluate the effects of an intervention to improve quality control in a manufacturing plant may not reveal important changes in performance if improvement in the percentage of correctly fabricated widgets from a baseline level of 92% to a range of 97 to 98% is the difference between unacceptable and acceptable (i.e., profitable) performance.

Threats to Measurement Accuracy and Reliability

The biggest threat to the accuracy and reliability of data in applied behavior analysis is human error. Unlike the experimental analysis of behavior, in which measurement is typically automated and conducted by machines, most investigations in applied behavior analysis use human observers to measure behavior.[3] Factors that contribute to human measurement error include poorly designed measurement systems, inadequate observer training, and expectations about what the data should look like.

Poorly Designed Measurement System

Unnecessarily cumbersome and difficult-to-use measurement systems create needless loss of accuracy and reliability. Collecting behavioral data in applied settings requires attention, keen judgment, and perseverance. The more taxing and difficult a measurement system is to use, the less likely an observer will be to consistently detect and record all instances of the target behavior. Simplifying the measurement system as much as possible minimizes measurement errors.

The complexity of measurement includes such variables as the number of individuals observed, the number of behaviors recorded, the duration of observation periods, and/or the duration of the observation intervals, all of which may affect the quality of measurement. For instance, observing several individuals is more complex than observing one person; recording several behaviors is more complex than recording a single behavior; using contiguous 5-second observation intervals with no time between intervals to record the results of the observation is more difficult than a system in which time is reserved for recording data.

Specific recommendations concerning reducing complexity depend on the specific nature of the study. However, when using time sampling measurements, applied behavior analysts can consider modifications such as decreasing the number of simultaneously observed individuals or behaviors, decreasing the duration of the observation sessions (e.g., from 30 minutes to 15 minutes), and increasing the duration of time intervals (e.g., from 5 to 10 seconds). Requiring more practice during observer training, establishing a higher criterion for mastery of the observational code, and providing more frequent feedback to observers may also reduce the possible negative effects of complex measurement.

Inadequate Observer Training

Careful attention must be paid to the selection and training of observers. Explicit and systematic training of observers is essential for the collection of trustworthy data. Observation and coding systems require observers to discriminate the occurrence and nonoccurrence of specific classes of behaviors or events against an often complex and dynamic background of other behaviors or events and to record their observations onto a data sheet. Observers must learn the definitions for each response class or event to be measured; a code or symbol notation system for each variable; a common set of recording procedures such as keystrokes or scan movements; and a method for correcting inadvertent handwritten, keystroke, or scan mistakes (e.g., writing a plus sign instead of a minus sign, hitting the F6 key instead of the F5 key, scanning an incorrect bar code).

Selecting Observers Carefully

Admittedly, applied researchers often scramble to find data collectors, but not all volunteers should be accepted into training. Potential observers should be interviewed to determine past experiences with observation and measurement activities, current schedule and upcoming

[3]We recommend using automatic data recording devices whenever possible. For example, to measure the amount of exercise by boys on stationary bicycles, DeLuca and Holborn (1992) used magnetic counters that automatically recorded the number of wheel revolutions.

commitments, work ethic and motivation, and overall social skills. The interview might include a pretest to determine current observation and skill levels. This can be accomplished by having potential observers watch short video clips of behaviors similar to what they may be asked to observe and noting their performance against a criterion.

Training Observers to an Objective Standard of Competency

Observer trainees should meet a specified criterion for recording before conducting observations in applied settings. During training, observers should practice recording numerous examples and nonexamples of the target behavior(s) and receive a critique and performance feedback. Observers should have numerous practice sessions before actual data collection. Training should continue until a predetermined criterion is achieved (e.g., 95% accuracy for two or three consecutive sessions). For example, in training observers to measure the completion of preventive maintenance tasks of heavy equipment by military personnel, Komaki (1998) required three consecutive sessions of at least 90% agreement with a true value.

Various methods can be used to train observers. These include sample vignettes, narrative descriptions, video sequences, role playing, and practice sessions in the environment in which actual data will be collected. Practice sessions in natural settings are especially beneficial because they allow both observers and participants to adapt to each other's presence and may reduce the reactive effects of the presence of observers on participants' behavior. The following steps are an example of a systematic approach for training observers.

Step 1 Trainees read the target behavior definitions and become familiar with data collection forms, procedures for recording their observations, and the proper use of any measurement or recording devices (e.g., tape recorders, stopwatches, laptops, PDAs, bar code scanners).

Step 2 Trainees practice recording simplified narrative descriptions of behavioral vignettes until they obtain 100% accuracy over a predetermined number of instances.

Step 3 Trainees practice recording longer, more complex narrative descriptions of behavioral vignettes until they obtain 100% accuracy for a predetermined number of episodes.

Step 4 Trainees practice observing and recording data from videotaped or role-played vignettes depicting the target behavior(s) at the same speed and complexity as they will occur in the natural environment. Training vignettes should be scripted and sequenced to provide trainees practice making increasingly difficult discriminations between the occurrence and nonoccurrence of the target behavior(s). Having trainees rescore the same series of vignettes a second time and comparing the reliability of their measures provides an assessment of the consistency with which the trainees are applying the measurement system. Trainees remain at this step until their data reach preestablished accuracy and reliability criteria. (If the study involved collecting data from natural permanent products such as compositions or academic worksheets, Steps 2 through 4 should provide trainees with practice scoring increasingly extensive and more difficult to score examples.)

Step 5 Practicing collecting data in the natural environment is the final training step of observer training. An experienced observer accompanies the trainee and simultaneously and independently measures the target behaviors. Each practice session ends with the trainee and experienced observer comparing their data sheets and discussing any questionable or heretofore unforeseen instances. Training continues until a preestablished criterion of agreement between the experienced observer and the trainee is achieved (e.g., at least 90% for three consecutive sessions).

Providing Ongoing Training to Minimize Observer Drift

Over the course of a study, observers sometimes alter, often unknowingly, the way they apply a measurement system. Called **observer drift,** these unintended changes in the way data are collected may produce measurement error. Observer drift usually entails a shift in the observer's interpretation of the definition of the target behavior from that used in training. Observer drift occurs when observers expand or compress the original definition of the target behavior. For example, observer drift might be responsible for the same behaviors by a child that were recorded by an observer as instances of noncompliance during the first week of a study being scored as instances of compliance during the study's final week. Observers are usually unaware of the drift in their measurement.

Observer drift can be minimized by occasional observer retraining or booster sessions throughout the investigation. Continued training provides the opportunity for observers to receive frequent feedback on the accuracy and reliability of measurement. Ongoing training can occur at regular, prescheduled intervals (e.g., every Friday morning) or randomly.

Unintended Influences on Observers

Ideally, data reported by observers have been influenced only by the actual occurrences and nonoccurrences of the target behavior(s) they have been trained to measure. In

reality, however, a variety of unintended and undesired influences on observers can threaten the accuracy and reliability of the data they report. Common causes of this type of measurement error include presuppositions an observer may hold about the expected outcomes of the data and an observer's awareness that others are measuring the same behavior.

Observer Expectations

Observer expectations that the target behavior should occur at a certain level under particular conditions, or change when a change in the environment has been made, pose a major threat to accurate measurement. For example, if an observer believes or predicts that a teacher's implementation of a token economy should decrease the frequency of inappropriate student behavior, she may record fewer inappropriate behaviors during the token reinforcement condition than she would have recorded otherwise without holding that expectation. Data influenced by an observer's expectations or efforts to obtain results that will please the researcher are characterized by measurement bias.

The surest way to minimize measurement bias caused by observer expectations is to use naive observers. A totally **naive observer** is a trained observer who is unaware of the study's purpose and/or the experimental conditions in effect during a given phase or observation period. Researchers should inform observer trainees that they will receive limited information about the study's purpose and why that is. However, maintaining observers' naiveté is often difficult and sometimes impossible.

When observers are aware of the purpose or hypothesized results of an investigation, measurement bias can be minimized by using target behavior definitions and recording procedures that will give a conservative picture of the behavior (e.g., whole-interval recording of on-task behavior with 10-second rather than 5-second intervals), frank and repeated discussion with observers about the importance of collecting accurate data, and frequent feedback to observers on the extent to which their data agree with true values or data obtained by observers who are naive. Observers should not receive feedback about the extent to which their data confirm or run counter to hypothesized results or treatment goals.

Observer Reactivity

Measurement error resulting from an observer's awareness that others are evaluating the data he reports is called **observer reactivity.** Like reactivity that may occur when participants are aware that their behavior is being observed, the behavior of observers (i.e., the data they record and report) can be influenced by the knowledge that others are evaluating the data. For example, knowing that the researcher or another observer is watching the same behavior at the same time, or will monitor the measurement through video- or audiotape later, may produce observer reactivity. If the observer anticipates that another observer will record the behavior in a certain way, his data may be influenced by what he anticipates the other observer may record.

Monitoring observers as unobtrusively as possible on an unpredictable schedule helps reduce observer reactivity. Separating multiple observers by distance or partition reduces the likelihood that their measures will be influenced by one another's during an observation. One-way mirrors in some research and clinical settings eliminate visual contact between the primary and secondary observers. If sessions are audiotaped or videotaped, the secondary observer can measure the behavior at a later time and the primary observer never has to come into contact with the secondary observer. In settings where one-way mirrors are not possible, and where audio- or videotaping may be intrusive, the secondary observer might begin measuring the behavior at a time unknown to the primary observer. For example, if the primary observer begins measuring behavior with the first interval, the secondary observer could start measuring behavior after 10 minutes have elapsed. The intervals used for comparisons would begin at the 10-minute mark, ignoring those intervals that the primary observer recorded beforehand.

Assessing the Accuracy and Reliability of Behavioral Measurement

After designing a measurement system that will produce a valid representation of the target behavior and training observers to use it in a manner that is likely to yield accurate and reliable data, the researcher's next measurement-related tasks are evaluating the extent to which the data are, in fact, accurate and reliable. Essentially, all procedures for assessing the accuracy and reliability of behavioral data entail some form of "measuring the measurement system."

Assessing the Accuracy of Measurement

Measurement is accurate when the observed values (i.e., the numbers obtained by measuring an event) match the true values of the event. The fundamental reason for determining the accuracy of data is obvious: No one wants to base research conclusions or make treatment decisions

on faulty data. More specifically, conducting accuracy assessments serves four interrelated purposes. First, it is important to determine early in an analysis whether the data are good enough to serve as the basis for making experimental or treatment decisions. The first person that the researcher or practitioner must try to convince that the data are accurate is herself. Second, accuracy assessments enable the discovery and correction of specific instances of measurement error. The two other approaches to assessing the quality of data to be discussed later in this chapter—reliability assessments and interobserver agreement—can alert the researcher to the likelihood of measurement errors, but neither approach identifies errors. Only the direct assessment of measurement accuracy allows practitioners or applied researchers to detect and correct faulty data.

A third reason for conducting accuracy assessments is to reveal consistent patterns of measurement error, which can lead to the overall improvement or **calibration** of the measurement system. When measurement error is consistent in direction and value, the data can be adjusted to compensate for the error. For example, knowing that John's bicycle computer reliably obtained a measure of 68 miles for a route with a true value of 58 miles led not only to the cyclists correcting the data in hand (in this case, confessing to one another and to their friend Lee that they had not ridden as many miles as previously claimed) but to their calibrating the measurement instrument so that future measures would be more accurate (in this case, adjusting the wheel circumference setting on John's bike computer).

Calibrating any measurement tool, whether it is a mechanical device or human observer, entails comparing the data obtained by the tool against a true value. The measure obtained by the Department of Transportation's wheel odometer served as the true value for calibrating John's bike computer. Calibration of a timing device such as a stopwatch or countdown timer could be made against a known standard: the "atomic clock."[4] If no differences are detected when comparing the timing device against the atomic clock, or if the differences are tolerable for the intended purposes of measurement, then calibration is satisfied. If significant differences are found, the timing device would need to be reset to the standard. We recommend frequent accuracy assessments in the beginning stages of an analysis. Then, if the assessments have produced high accuracy, less frequent assessments can be conducted to check the calibration of the recorders.

A fourth reason for conducting accuracy assessments is to assure consumers that the data are accurate. Including the results of accuracy assessments in research reports helps readers judge the trustworthiness of the data being offered for interpretation.

Establishing True Values

"There is only one way to assess the accuracy of a set of measures—by comparing observed values to true values. The comparison is relatively easy; the challenge is often obtaining measures of behavior that can legitimately be considered true values" (Johnston & Pennypacker, 1993a, p. 138). As defined previously, a *true value* is a measure obtained by procedures that are independent of and different from the procedures that produced the data being evaluated and for which the researcher has taken "special or extraordinary precautions to ensure that all possible sources of error have been avoided or removed" (p. 136).

True values for some behaviors are evident and universally accepted. For example, obtaining the true values of correct responses in academic areas such math and spelling is straightforward. The correct response to the arithmetic problem 2 + 2 = ? has a true value of 4, and the *Oxford English Dictionary* is a source of true values for assessing the accuracy of measuring the spelling of English words.[5] Although not universal, true values for many socially significant behaviors of interest to applied researchers and practitioners can be established conditionally on local context. For example, the correct response to the question "Name the three starches recommended as thickeners for pan gravy" on a quiz given to students in a culinary school has no universal true value. Nevertheless, a true value relevant to the students taking the quiz can be found in the instructor's course materials.

True values for each of the preceding examples were obtained through sources independent of the measures to be evaluated. Establishing true values for many behaviors studied by applied behavior analysts is difficult because the process for determining a true value must be different from the measurement procedures used to obtain the data one wishes to compare to the true value. For example, determining true values for occurrences of a behavior such as cooperative play between children is difficult because the only way to attach any values to the behavior is to measure it with the same observation procedures used to produce the data in the first place.

It can be easy to mistake true values as values that only appear to be true values. For example, suppose that

[4]The official time in the United States can be accessed through the National Bureau of Standards and the United States Naval Observatory atomic clock (actually 63 atomic clocks are averaged to determine official time): http://tycho.usno.navy.mil/what1.html. The atomic clock is accurate to 1 billionth of a second per day, or 1 second per 6 million years!

[5]The preferred spelling of a word may change (e.g., *judgement* becomes *judgment*), but in such cases a new true value is established.

four well-trained and experienced observers view a videotape of teacher and student interactions. Their task is to identify the true value of all instances of teacher praise contingent on academic accomplishments. Each observer views the tape independently and counts all occurrences of contingent teacher praise. After recording their respective observations, the four observers share their measurements, discuss disagreements, and suggest reasons for the disagreements. The observers independently record contingent praise a second time. Once again they share and discuss their results. After repeating the recording and sharing process several times, all observers agree that they have recorded every instance of teacher praise. However, the observers did not produce a *true value* of teacher praise for two reasons: (1) The observers could not calibrate their measurement of teacher praise to an independent standard of teacher praise, and (2) the process used to identify all instances of teacher praise may be biased (e.g., one of the observers may have convinced the others that her measures represented the true value). When true values cannot be established, researchers must rely on reliability assessments and measures of interobserver agreement to evaluate the quality of their data.

Accuracy Assessment Procedures

Determining the accuracy of measurement is a straightforward process of calculating the correspondence of each measure, or datum, assessed to its true value. For example, a researcher or practitioner assessing the accuracy of the score for a student's performance on a 30-word spelling test reported by a grader would compare the grader's scoring of each word on the test with the true value for that word found in a dictionary. Each word on the test that matched the correct letter sequence (i.e., orthography) provided by the dictionary and was marked correct by the grader would be an accurate measure by the grader, as would each word marked incorrect by the grader that did not match the dictionary's spelling. If the original grader's scoring of 29 of the test's 30 words corresponded to the true values for those words, the grader's measure would be 96.7% accurate.

Although an individual researcher or practitioner can assess the accuracy of the data she has collected, multiple independent observers are often used. Brown, Dunne, and Cooper (1996) described the procedures they used to assess the accuracy of measurement in a study of oral reading comprehension as follows:

> An independent observer reviewed one student's audiotape of the delayed one-minute oral retell each day to assess our accuracy of measurement, providing an assessment of the extent that our counts of delayed retells approximated the true value of the audio-taped correct and incorrect retells. The independent observer randomly selected each day's audiotape by drawing a student's name from a hat, then listened to the tape and scored correct and incorrect retells using the same definitions as the teacher. Observer scores were compared to teacher scores. If there was a discrepancy between these scores, the observer and the teacher reviewed the tape (i.e., the true value) together to identify the source of the discrepancy and corrected the counting error on the data sheet and the Standard Celeration Chart. The observer also used a stopwatch to time the duration of the audiotape to ensure accuracy of the timings. We planned to have the teacher re-time the presentation or retell and recalculate the frequency per minute for each timing discrepancy of more than 5 seconds. All timings, however, met the 5-second accuracy definition. (p. 392)

Reporting Accuracy Assessments

In addition to describing procedures used to assess the accuracy of the data, researchers should report the number and percentage of measures that were checked for accuracy, the degree of accuracy found, the extent of measurement error detected, and whether those measurement errors were corrected in the data. Brown and colleagues (1996) used the following narrative to report the results of their accuracy assessment:

> The independent observer and the teacher achieved 100% agreement on 23 of the 37 sessions checked. The teacher and the observer reviewed the tape together to identify the source of measurement errors for the 14 sessions containing measurement discrepancies and corrected the measurement errors. Accurate data from the 37 sessions rechecked were then displayed on the Standard Celeration Charts. The magnitude of the measurement errors was very small, often a difference of 1 to 3 discrepancies. (p. 392)

A full description and reporting of the results of accuracy assessment helps readers of the study evaluate the accuracy of all of the data included in the report. For example, suppose a researcher reported that she conducted accuracy checks on a randomly selected 20% of the data, found those measures to be 97% accurate with the 3% error being nonbiased, and corrected the assessed data as needed. A reader of the study would know that 20% of the data are 100% accurate and be fairly confident that the remaining 80% of the data (i.e., all of the measures that were not checked for accuracy) is 97% accurate.

Assessing the Reliability of Measurement

Measurement is reliable when it yields the same values across repeated measures of the same event. Reliability is established when the same observer measures the same

data set repeatedly from archived response products such as audiovisual products and other forms of permanent products. The more frequently a consistent pattern of observation is produced, the more reliable the measurement (Thompson et al., 2000). Conversely, if similar observed values are not achieved with repeated observations, the data are considered unreliable. This leads to a concern about accuracy, which is the primary indicator of quality measurement.

But, as we have pointed out repeatedly, reliable data are not necessarily accurate data. As the three bicyclists discovered, totally reliable (i.e., consistent) measurement may be totally wrong. Relying on the reliability of measurement as the basis for determining the accuracy of measurement would be, as the philosopher Wittgenstein (1953) noted, "As if someone were to buy several copies of the morning paper to assure himself that what it said was true" (p. 94).

In many research studies and most practical applications, however, checking the accuracy of every measure is not possible or feasible. In other cases, true values for measures of the target behavior may be difficult to establish. When confirming the accuracy of each datum is not possible or practical, or when true values are not available, knowing that a measurement system has been applied with a high degree of consistency contributes to confidence in the overall trustworthiness of the data. Although high reliability cannot confirm high accuracy, discovering a low level of reliability signals that the data are then suspect enough to be disregarded until problems in the measurement system can be determined and repaired.

Assessing the reliability of behavioral measurement requires either a natural or contrived permanent product so the observer can remeasure the same events. For example, reliability of measurement of variables such as the number of adjectives or action verbs in students' essays could be accomplished by having an observer rescore essays. Reliability of measurement of the number and type of response prompts and feedback statements by parents to their children at the family dinner table could be assessed by having an observer replay and rescore videotapes of the family's mealtime and compare the data obtained from the two measurements.

Observers should not remeasure the same permanent product soon after measuring it the first time. Doing so might result in the measures from the second scoring being influenced by what the observer remembered from the initial scoring. To avoid such unwanted influence, a researcher can insert several previously scored essays or videotapes randomly into the sequence of "new data" being recorded by observers.

Using Interobserver Agreement to Assess Behavioral Measurement

Interobserver agreement is the most commonly used indicator of measurement quality in applied behavior analysis. **Interobserver agreement (IOA)** refers to the degree to which two or more independent observers report the same observed values after measuring the same events. There are numerous techniques for calculating IOA, each of which provides a somewhat different view of the extent and nature of agreement and disagreement between observers (e.g., Hartmann, 1977; Hawkins & Dotson, 1975; Page & Iwata, 1986; Poling, Methot, & LeSage, 1995; Repp, Dietz, Boles, Dietz, & Repp, 1976).

Benefits and Uses of IOA

Obtaining and reporting interobserver agreement serves four distinct purposes. First, a certain level of IOA can be used as a basis for determining the competence of new observers. As noted earlier, a high degree of agreement between a newly trained observer and an experienced observer provides an objective index of the extent to which the new observer is measuring the behavior in the same way as experienced observers.

Second, systematic assessment of IOA over the course of a study can detect observer drift. When observers who obtained the same, or nearly the same, observed values when measuring the same behavioral events at the beginning of a study (i.e., IOA was high) obtain different measures of the same events later in the study (i.e., IOA is now low), one of the observers may be using a definition of the target behavior that has drifted. Deteriorating IOA assessments cannot indicate with assurance which of the observer's data are being influenced by drift (or any other reason for disagreement), but the information reveals the need for further evaluation of the data and/or for retraining and calibration of the observers.

Third, knowing that two or more observers consistently obtained similar data increases confidence that the definition of the target behavior was clear and unambiguous and the measurement code and system not too difficult. Fourth, for studies that employ multiple observers as data collectors, consistently high levels of IOA increase confidence that variability in the data is not a function of which observer(s) happened to be on duty for any given session, and therefore that changes in the data more likely reflect actual changes in the behavior.

The first two reasons for assessing IOA are proactive: They help researchers determine and describe the degree to which observers have met training criteria and detect possible drift in observers' use of the measurement

system. The second two purposes or benefits of IOA are as summative descriptors of the consistency of measurement across observers. By reporting the results of IOA assessments, researchers enable consumers to judge the relative **believability** of the data as trustworthy and deserving of interpretation.

Requisites for Obtaining Valid IOA Measures

A valid assessment of IOA depends on three equally important criteria. Although these criteria are perhaps obvious, it is nonetheless important to make them explicit. Two observers (usually two, but may be more) must (a) use the same observation code and measurement system, (b) observe and measure the same participant(s) and events, and (c) observe and record the behavior independent of any influence from one other.

Observers Must Use the Same Measurement System

Interobserver agreement assessments conducted for any of the four previously stated reasons require observers to use the same definitions of the target behavior, observation procedures and codes, and measurement devices. Beyond using the same measurement system, all observers participating in IOA measures used to assess the believability of data (as opposed to evaluating the observer trainees' performance) should have received identical training with the measurement system and achieved the same level of competence in using it.

Observers Must Measure the Same Events

The observers must be able to observe the same subject(s) at precisely the same observation intervals and periods. IOA for data obtained by real-time measurement requires that both observers be in the setting simultaneously. Real-time observers must be positioned such that each has a similar view of the subject(s) and environment. Two observers sitting on opposite sides of a classroom, for example, might obtain different measures because the different vantage points enable only one observer to see or hear some occurrences of the target behavior.

Observers must begin and end the observation period at precisely the same time. Even a difference of a few seconds between observers may produce significant measurement disagreements. To remedy this situation, the timing devices could be started simultaneously and outside the observation setting, but before data collection begins, with the understanding that the data collection would actually start at a prearranged time (e.g., exactly at the beginning of the fifth minute). Alternatively, but less desirably, one observer could signal the other at the exact moment the observation is to begin.

A common and effective procedure is for both observers to listen by earphones to an audiotape of prerecorded cues signaling the beginning and end of each observation interval (see Chapter 4). An inexpensive splitter device that enables two earphones to be plugged into the same tape recorder allows observers to receive simultaneous cues unobtrusively and without depending on one another.

When assessing IOA for data obtained from permanent products, the two observers do not need to measure the behavior simultaneously. For example, the observers could each watch and record data from the same video- or audiotape at different times. Procedures must be in place, however, to ensure that each observer watched or listened to the same tapes and that they started and stopped their independent observations at precisely the same point(s) on the tapes. Ensuring that two observers measure the same events when the target behavior produces natural permanent products, such as completed academic assignments or widgets manufactured, would include procedures such as clearly marking the session number, date, condition, and subject's name on the product and guarding the response products to ensure that they are not disturbed until the second observer has obtained his measure.

Observers Must Be Independent

The third essential ingredient for valid IOA assessment is ensuring that neither observer is influenced by the other's measurements. Procedures must be in place to guarantee each observer's independence. For example, observers conducting real-time measurement of behavior "must be situated so that they can neither see nor hear when the other observes and records a response" (Johnston & Pennypacker, 1993a, p. 147). Observers must not be seated or positioned so closely to one another that either observer can detect or be influenced by the other observer's recordings.

Giving the second observer academic worksheets or written assignments that have already been marked by another observer would violate the observers' independence. To maintain independence, the second observer must score photocopies of unadulterated and unmarked worksheets or assignments as completed by the subjects.

Methods for Calculating IOA

There are numerous methods for calculating IOA, each of which provides a somewhat different view of the extent and nature of agreement and disagreement between observers (e.g., Hartmann, 1977; Hawkins & Dotson, 1975;

Page & Iwata, 1986; Poling, Methot, & LeSage, 1995; Repp, Dietz, Boles, Dietz, & Repp, 1976). The following explanation of different IOA formats is organized by the three major methods for measuring behavioral data described in Chapter 4: event recording, timing, and interval recording or time sampling. Although other statistics are sometimes used, the percentage of agreement between observers is by far the most common convention for reporting IOA in applied behavior analysis.[6] Therefore, we have provided the formula for calculating a percentage of agreement for each type of IOA.

IOA for Data Obtained by Event Recording

The various methods for calculating interobserver agreement for data obtained by event recording are based on comparing (a) the total count recorded by each observer per measurement period, (b) the counts tallied by each observer during each of a series of smaller intervals of time within the measurement period, or (c) each observer's count of 1 or 0 on a trial-by-trial basis.

Total Count IOA.[7] The simplest and crudest indicator of IOA for event recording data compares the total count recorded by each observer per measurement period. **Total count IOA** is expressed as a percentage of agreement between the total number of responses recorded by two observers and is calculated by dividing the smaller of the counts by the larger count and multiplying by 100, as shown by this formula:

$$\frac{\text{Smaller count}}{\text{Larger count}} * 100 = \text{total count IOA \%}$$

For example, suppose that a child care worker in a residential setting recorded that 9-year-old Mitchell used profane language 10 times during a 30-minute observation period and that a second observer recorded that Mitchell swore 9 times during that same period. The total count IOA for the observation period would be 90% (i.e., $9 \times 10 \div 100 = 90\%$).

Great caution must be used in interpreting total count IOA because a high degree of agreement provides no assurance that the two observers recorded the same instances of behavior. For example, the following is one of the countless ways that the data reported by the two observers who measured Mitchell's use of profane language may not represent anywhere close to 90% agreement that they measured the same behaviors. The child care worker could have recorded all 10 occurrences of profane language on her data sheet during the first 15 minutes of the 30-minute observation period, a time when the second observer recorded just 4 of the 9 total responses he reported.

Mean Count-per-Interval IOA. The likelihood that significant agreement between observers' count data means they measured the same events can be increased by (a) dividing the total observation period into a series of smaller counting times, (b) having the observers record the number of occurrences of the behavior within each interval, (c) calculating the agreement between the two observers' counts within each interval, and (d) using the agreements per interval as the basis for calculating the IOA for the total observation period. The hypothetical data shown in Figure 5.2 will be used to illustrate two methods for calculating count-per-interval IOA: mean count-per-interval and exact count-per-interval. During a 30-minute observation period, two observers independently tallied the number of times each witnessed an instance of a target behavior during each of six 5-minute intervals.

Even though each observer recorded a total of 15 responses within the 30-minute period, their data sheets reveal a high degree of disagreement within the observation period. Although the total count IOA for the entire observation period was 100%, agreement between the two observers within each 5-minute interval ranged from 0% to 100%, yielding a mean count-per-interval IOA of 65.3%.

Mean count-per-interval IOA is calculated by this formula:

$$\frac{\text{Int 1 IOA} + \text{Int 2 IOA} + \text{Int N IOA}}{\text{n intervals}} * 100 = \text{mean count-per-interval IOA \%}$$

Exact Count-per-Interval IOA. The most stringent description of IOA for most data sets obtained by event recording is obtained by computing the **exact count-per-interval IOA**—the percentage of total intervals in which two observers recorded the same count. The two observers whose data are shown in Figure 5.2 recorded

[6] IOA can be calculated by product-moment correlations, which range from +1.0 to −1.0. However, expressing IOA by correlation coefficients has two major weaknesses: (a) High coefficients can be achieved if one observer consistently records more occurrences of the behavior than the other, and (b) correlation coefficients provide no assurance that the observers agreed on the occurrence of any given instance of behavior (Poling et al., 1995). Hartmann (1977) described the use of *kappa (k)* as an measure of IOA. The *k* statistic was developed by Cohen (1960) as a procedure for determining the proportion of agreements between observers that would be expected as a result of chance. However, the *k* statistic is seldom reported in the behavior analysis literature.

[7] Multiple terms are used in the applied behavior analysis literature for the same methods of calculating IOA, and the same terms are sometimes used with different meanings. We believe the IOA terms used here represent the discipline's most used conventions. In an effort to point out and preserve some meaningful distinctions among variations of IOA measures, we have introduced several terms.

Figure 5.2 Two methods for computing interobserver agreement (IOA) for event recording data tallied within smaller time intervals.

Interval (Time)	Observer 1	Observer 2	IOA per interval
1 (1:00–1:05)	///	//	2/3 = 67%
2 (1:05–1:10)	///	///	3/3 = 100%
3 (1:10–1:15)	/	//	1/2 = 50%
4 (1:15–1:20)	////	///	3/4 = 75%
5 (1:20–1:25)	0	/	0/1 = 0%
6 (1:25–1:30)	////	////	4/4 = 100%
	Total count = 15	**Total count = 15**	**Mean count-per-interval IOA = 65.3% Exact count-per-interval IOA = 33%**

the same number of responses in just two of the six intervals, an exact count-per-interval IOA of 33%.

The following formula is used to calculate exact count-per-interval IOA:

$$\frac{\text{Number of intervals of 100\% IOA}}{n \text{ intervals}} \times 100 = \text{exact count-per-interval IOA \%}$$

Trial-by-Trial IOA. The agreement between two observers who measured the occurrence or nonoccurrence of discrete trial behaviors for which the count for each trial, or response opportunity, can only be 0 or 1 can be calculated by comparing the observers' total counts or by comparing their counts on a trial-by-trial basis. Calculating total count IOA for discrete trial data uses the same formula as total count IOA for free operant data: The smaller of the two counts reported by the observers is divided by the larger count and multiplied by 100, but in this case the number of trials for which each observer recorded the occurrence of the behavior is the count. Suppose, for example, that a researcher and a second observer independently measured the occurrence or nonoccurrence of a child's smiling behavior during each of 20 trials that the researcher showed the child a funny picture. The two observers compare data sheets at the end of the session and discover that they recorded smiles on 14 and 15 trials, respectively. The total count IOA for the session is 93% (i.e., 14 ÷ 15 × 100 = 93.3%), which might lead an inexperienced researcher to conclude that the target behavior has been well defined and is being measured with consistency by both observers. Those conclusions, however, would not be warranted.

Total count IOA of discrete trial data is subject to the same limitations as total count IOA of free operant data: It tends to overestimate the extent of actual agreement and does not indicate how many responses, or which responses, trials, or items, posed agreement problems. Comparing the two observers' counts of 14 and 15 trials suggests that they disagreed on the occurrence of smiling on only 1 of 20 trials. However, it is possible that any of the 6 trials scored as "no smile" by the experimenter was scored as a "smile" trial by the second observer and that any of the 5 trials recorded by the second observer as "no smile" was recorded as a "smile" by the experimenter. Thus, the total count IOA of 93% may vastly overestimate the actual consistency with which the two observers measured the child's behavior during the session.

A more conservative and meaningful index of interobserver agreement for discrete trial data is **trial-by-trial IOA,** which is calculated by the following formula:

$$\frac{\text{Number of trials (items) agreement}}{\text{Total number of trials (items)}} \times 100 = \text{trial-by-trial IOA \%}$$

The trial-by-trial IOA for the two observers' smiling data, if calculated with the worst possible degree of agreement from the previous example—that is, if all 6 trials that the primary observer scored as "no smile" were recorded as "smile" trials by the second observer and all 5 trials marked by the second observer as "no smile" were recorded as "smile" trials by the experimenter—would be 45% (i.e., 9 trials scored in agreement divided by 20 trials × 100).

IOA for Data Obtained by Timing

Interobserver agreement for data obtained by timing duration, response latency, or interresponse time (IRT) is obtained and calculated in essentially the same way as it

is for event recording data. Two observers independently time the duration, latency, or IRT of the target behavior, and IOA is based on comparing either the total time obtained by each observer for the session or the times recorded by each observer per occurrence of the behavior (for duration measures) or per response (for latency and IRT measures).

Total Duration IOA. **Total duration IOA** is computed by dividing the shorter of the two durations reported by the observers by the longer duration and multiplying by 100.

$$\frac{\text{Shorter duration}}{\text{Longer duration}} \times 100 = \text{total duration IOA \%}$$

As with total count IOA for event recording data, high total duration IOA provides no assurance that the observers recorded the same durations for the same occurrences of behavior. This is because a significant degree of disagreement between the observers' timings of individual responses may be canceled out in the sum. For example, suppose two observers recorded the following durations in seconds for five occurrences of a behavior:

	R1	R2	R3	R4	R5
Observer 1: (total duration = 90 seconds)	35	15	9	14	17
Observer 2: (total duration = 85 seconds)	29	21	7	14	14

Total duration IOA for these data is a perhaps comforting 94% (i.e., 85 ÷ 90 × 100 = 94.4%). However, the two observers obtained the same duration for only one of the five responses, and their timings of specific responses varied by as much as 6 seconds. While recognizing this limitation of total duration IOA, when total duration is being recorded and analyzed as a dependent variable, reporting total duration IOA is appropriate. When possible, total duration IOA should be supplemented with mean duration-per-occurrence IOA, which is described next.

Mean Duration-per-Occurrence IOA. Mean duration-per-occurrence IOA should be calculated for duration per occurrence data, and it is a more conservative and usually more meaningful assessment of IOA for total duration data. The formula for calculating **mean duration-per-occurrence IOA** is similar to the one used to determine mean count-per-interval IOA:

$$\frac{\text{Dur IOA R1} + \text{Dur IOA R2} + \text{Dur IOA R}n}{n \text{ responses with Dur IOA}} \times 100 = \text{mean duration-per-interval IOA \%}$$

Using this formula to calculate the mean duration-per-occurrence IOA for the two observers' timing data of the five responses just presented would entail the following steps:

1. Calculate duration per occurrence IOA for each response: R1, 29 ÷ 35 = .83; R2, 15 ÷ 21 = .71; R3, 7 ÷ 9 = .78; R4, 14 ÷ 14 = 1.0; and R5, 14 ÷ 17 = .82
2. Add the individual IOA percentages for each occurrence: .83 + .71 + .78 + 1.00 + .82 = 4.14
3. Divide the sum of the individual IOAs per occurrence by the total number of responses for which two observers measured duration: 4.14 ÷ 5 = .828
4. Multiply by 100 and round to the nearest whole number: .828 × 100 = 83%

This basic formula is also used to compute the *mean latency-per-response IOA* or *mean IRT-per-response IOA* for latency and IRT data. An observer's timings of latencies or IRTs in a session should never be added and the total time compared to a similar total time obtained by another observer as the basis for calculating IOA for latency and IRT measures.

In addition to reporting mean agreement per occurrence, IOA assessment for timing data can be enhanced with information about the range of differences between observers' timings and the percentage of responses for which the two observers each obtained measures within a certain range of error. For example: Mean duration-per-occurrence IOA for Temple's compliance was 87% (range across responses, 63 to 100%), and 96% of all timings obtained by the second observer were within +/–2 seconds of the primary observer's measures.

IOA for Data Obtained by Interval Recording/Time Sampling

Three techniques commonly used by applied behavior analysts to calculate IOA for interval data are interval-by-interval IOA, scored-interval IOA, and unscored-interval IOA.

Interval-by-Interval IOA. When using an interval-by-interval IOA (sometimes referred to as the *point-by-point* and *total interval* method), the primary observer's record for each interval is matched to the secondary observer's record for the same interval. The formula for calculating **interval-by-interval IOA** is as follows:

$$\frac{\text{Number of intervals agreed}}{\text{Number of intervals agreed} + \text{number of intervals disagreed}} \times 100 = \text{interval-by-interval IOA \%}$$

Figure 5.3 When calculating interval-by-interval IOA, the number of intervals in which both observers agreed on the occurrence or the nonoccurrence of the behavior (shaded intervals) is divided by the total number of observation intervals. Interval-by-interval IOA for the data shown here is 70% (7/10).

Interval-by-Interval IOA

Interval no. →	1	2	3	4	5	6	7	8	9	10
Observer 1	X	X	X	0	X	X	0	X	X	0
Observer 2	0	X	X	0	X	0	0	0	X	0

X = behavior was recorded as occurring during interval
0 = behavior was recorded as not occurring during interval

The hypothetical data in Figure 5.3 show the interval-by-interval method for calculating IOA based on the record of two observers who recorded the occurrence (X) and nonoccurrence (0) of behavior in each of 10 observation intervals. The observers' data sheets show that they agreed on the occurrence or the nonoccurrence of the behavior for seven intervals (Intervals 2, 3, 4, 5, 7, 9, and 10). Interval-by-interval IOA for this data set is 70% (i.e., $7 \div [7 + 3] \times 100 = 70\%$).

Interval-by-interval IOA is likely to overestimate the actual agreement between observers measuring behaviors that occur at very low or very high rates. This is because interval-by-interval IOA is subject to random or accidental agreement between observers. For example, with a behavior whose actual frequency of occurrence is only about 1 or 2 intervals per 10 observation intervals, even a poorly trained and unreliable observer who misses some of the few occurrences of the behavior and mistakenly records the behavior as occurring in some intervals when the behavior did not occur is likely to mark most intervals as nonoccurrences. As a result of this chance agreement, interval-by-interval IOA is likely to be quite high. Two IOA methods that minimize the effects of chance agreements for interval data on behaviors that occur at very low or very high rates are scored-interval IOA and unscored-interval IOA (Hawkins & Dotson, 1975).

Scored-Interval IOA. Only those intervals in which either or both observers recorded the *occurrence of the target behavior* are used in calculating **scored-interval IOA.** An agreement is counted when both observers recorded that the behavior occurred in the same interval, and each interval in which one observer recorded the occurrence of the behavior and the other recorded its nonoccurrence is counted as a disagreement. For example, for the data shown in Figure 5.4, only Intervals 1, 3, and 9 would be used in calculating scored-interval IOA. Intervals 2, 4, 5, 6, 7, 8, and 10 would be ignored because both observers recorded that the behavior did not occur in those intervals. Because the two observers agreed that the behavior occurred in only one (Interval 3) of the three scored intervals, the scored-interval IOA measure is 33% (1 interval of agreement divided by the sum of 1 interval of agreement plus 2 intervals of disagreement $\times$ 100 = 33%).

Figure 5.4 Scored-interval IOA is calculated using only those intervals in which either observer recorded the occurrence of the behavior (shaded intervals). Scored-interval IOA for the data shown here is 33% (1/3).

Scored-Interval IOA

Interval no. →	1	2	3	4	5	6	7	8	9	10
Observer 1	X	0	X	0	0	0	0	0	0	0
Observer 2	0	0	X	0	0	0	0	0	X	0

X = behavior was recorded as occurring during interval
0 = behavior was recorded as not occurring during interval

For behaviors that occur at low rates, scored-interval IOA is a more conservative measure of agreement than interval-by-interval IOA. This is because scored-interval IOA ignores the intervals in which agreement by chance is highly likely. For example, using the interval-by-interval method for calculating IOA for the data in Figure 5.4 would yield an agreement of 80%. To avoid overinflated and possibly misleading IOA measures, we recommend using scored-interval interobserver agreement for behaviors that occur at frequencies of approximately 30% or fewer intervals.

Unscored-Interval IOA. Only intervals in which either or both observers recorded the *nonoccurrence of the target behavior* are considered when calculating **unscored-interval IOA.** An agreement is counted when both observers recorded the nonoccurrence of the behavior in the same interval, and each interval in which one observer recorded the nonoccurrence of the behavior and the other recorded its occurrence is counted as a disagreement. For example, only Intervals 1, 4, 7, and 10 would be used in calculating the unscored-interval IOA for the data in Figure 5.5 because at least one observer recorded the nonoccurrence of the behavior in each of those intervals. The two observers agreed that the behavior did not occur in Intervals 4 and 7. Therefore, the unscored-interval IOA in this example is 50% (2 intervals of agreement divided by the sum of 2 intervals of agreement plus 2 intervals of disagreement × 100 = 50%).

For behaviors that occur at relatively high rates, unscored-interval IOA provides a more stringent assessment of interobserver agreement than does interval-by-interval IOA. To avoid overinflated and possibly misleading IOA measures, we recommend using unscored-interval interobserver agreement for behaviors that occur at frequencies of approximately 70% or more of intervals.

Considerations in Selecting, Obtaining, and Reporting Interobserver Agreement

The guidelines and recommendations that follow are organized under a series of questions concerning the use of interobserver agreement to evaluate the quality of behavioral measurement.

How Often and When Should IOA Be Obtained?

Interobserver agreement should be assessed during each condition and phase of a study and be distributed across days of the week, times of day, settings, and observers. Scheduling IOA assessments in this manner ensures that the results will provide a representative (i.e., valid) picture of all data obtained in a study. Current practice and recommendations by authors of behavioral research methods texts suggest that IOA be obtained for a minimum of 20% of a study's sessions, and preferably between 25% and 33% of sessions (Kennedy, 2005; Poling et al., 1995). In general, studies using data obtained via real-time measurement will have IOA assessed for a higher percentage of sessions than studies with data obtained from permanent products.

The frequency with which data should be assessed via interobserver agreement will vary depending on the complexity of the measurement code, the number and experience of observers, the number of conditions and phases, and the results of the IOA assessments themselves. More frequent IOA assessments are expected in studies that involve complex or new measurement systems, inexperienced observers, and numerous conditions and phases. If appropriately conservative methods for obtaining and calculating IOA reveal high levels of agreement early in a study, the number and proportion of sessions in which IOA is assessed may decrease as the study progresses. For instance, IOA assessment might be conducted in each

Figure 5.5 Unscored-interval IOA is calculated using only those intervals in which either observer recorded the nonoccurrence of the behavior (shaded intervals). Unscored interval IOA for the data shown here is 50% (2/4).

Unscored-Interval IOA

Interval no. →	1	2	3	4	5	6	7	8	9	10
Observer 1	X	X	X	0	X	X	0	X	X	0
Observer 2	0	X	X	0	X	X	0	X	X	X

X = behavior was recorded as occurring during interval

0 = behavior was recorded as not occurring during interval

session at the beginning of an analysis, and then reduced to a schedule of once per four or five sessions.

For What Variables Should IOA Be Obtained and Reported?

In general, researchers should obtain and report IOA at the same levels at which they report and discuss the results of their study. For example, a researcher analyzing the relative effects of two treatment conditions on two behaviors of four participants in two settings should report IOA outcomes on both behaviors for each participant separated by treatment condition and setting. This would enable consumers of the research to judge the relative believability of the data within each component of the experiment.

Which Method of Calculating IOA Should Be Used?

More stringent and conservative methods of calculating IOA should be used over methods that are likely to overestimate actual agreement as a result of chance. With event recording data used to evaluate the accuracy of performance, we recommend reporting overall IOA on a trial-by-trial or item-by-item basis, perhaps supplemented with separate IOA calculations for correct responses and incorrect responses. For data obtained by interval or time sampling measurement, we recommend supplementing interval-by-interval IOA with scored-interval IOA or unscored-interval IOA depending on the relative frequency of the behavior. In situations in which the primary observer scores the target behavior as occurring in approximately 30% or fewer intervals, scored-interval IOA provides a conservative supplement to interval-by-interval IOA. Conversely, when the primary observer scores the target behavior as occurring in approximately 70% or more of the intervals, unscored-interval IOA should supplement interval-by-interval IOA. If the rate at which the target behavior occurs changes from very low to very high, or from very high to very low, across conditions or phases of a study, reporting both unscored-interval and scored-interval IOA may be warranted.

If in doubt about which form of IOA to report, calculating and presenting several variations will help readers make their own judgments regarding the believability of the data. However, if the acceptance of the data for interpretation or decision making rests on which formula for calculating IOA is chosen, serious concerns about the data's trustworthiness exist that must be addressed.

What Are Acceptable Levels of IOA?

Carefully collected and conservatively computed IOA assessments increasingly enhance the believability of a data set as agreement approaches 100%. The usual convention in applied behavior analysis is to expect independent observers to achieve a mean of no less than 80% agreement when using observational recording. However, as Kennedy (2005) pointed out, "There is no scientific justification for why 80% is necessary, only a long history of researchers using this percentage as a benchmark of acceptability and being successful in their research activities" (p. 120).

Miller (1997) recommended that IOA should be 90% or greater for an established measure and at least 80% for a new variable. Various factors at work in a given situation may make an 80% or 90% criterion too low or too high. Interobserver agreement of 90% on the number of words contained in student compositions should raise serious questions about the trustworthiness of the data. IOA near 100% is needed to enhance the believability of count data obtained from permanent products. However, some analysts might accept data with a mean IOA as low as 75% for the simultaneous measurement of multiple behaviors by several subjects in a complex environment, especially if it is based on a sufficient number of individual IOA assessments with a small range (e.g., 73 to 80%).

The degree of behavior change revealed by the data should also be considered when determining an acceptable level of interobserver agreement. When behavior change from one condition to another is small, the variability in the data might represent inconsistent observation more than actual change in the behavior. Therefore, the smaller the change in behavior across conditions, the higher the criterion should be for an acceptable IOA percentage (Kennedy, 2005).

How Should IOA Be Reported?

IOA scores can be reported in narrative, table, and graphic form. Whichever format is chosen, it is important to note how, when, and how often interobserver agreement was assessed.

Narrative Description. The most common approach for reporting IOA is a simple narrative description of the mean and range of agreement percentages. For example, Craft, Alber, and Heward (1998) described the methods and results of IOA assessments in a study in which four dependent variables were measured as follows:

> *Student recruiting and teacher praise.* A second observer was present for 12 (30%) of the study's 40

sessions. The two observers independently and simultaneously observed the 4 students, recording the number of recruiting responses they emitted and teacher praise they received. Descriptive narrative notes recorded by the observers enabled each recruiting episode to be identified for agreement purposes. Interobserver agreement was calculated on an episode-by-episode basis by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100%. Agreement for frequency of student recruiting ranged across students from 88.2% to 100%; agreement for frequency of recruited teacher praise was 100% for all 4 students; agreement for frequency of nonrecruited teacher praise ranged from 93.3% to 100%.

Academic work completion and accuracy. A second observer independently recorded each student's work completion and accuracy for 10 (25%) sessions. Interobserver agreement for both completion and accuracy on the spelling worksheets was 100% for all 4 students.

Table. An example of reporting interobserver agreement outcomes in table format is shown in Table 5.1. Krantz and McClannahan (1998) reported the range and mean IOA computed for three types of social interactions by three children across each experimental condition.

Graphic Display. Interobserver agreement can be represented visually by plotting the measures obtained by the secondary observer on a graph of the primary observer's data as shown in Figure 5.6. Looking at both observers' data on the same graph reveals the extent of agreement between the observers and the existence of observer drift or bias. The absence of observer drift is suggested in the hypothetical study shown in Figure 5.6 because the secondary observer's measures changed in concert with the primary observer's measures. Although the two observers obtained the same measure on only 2 of the 10 sessions in which IOA was assessed (Sessions 3 and 8), the fact that neither observer consistently reported measures that were higher or lower than the other suggests the absence of observer bias. An absence of bias is usually indicated by a random pattern of overestimation and underestimation. In addition to revealing observer drift and bias, a third way that graphically displaying IOA assessments can enhance the believability of measurement is illustrated by the

Table 5.1 Interobserver Agreement Results for Each Dependent Variable by Participant and Experimental Condition

Range and Mean Percentage Interobserver Agreement on Scripted Interaction, Elaborations, and Unscripted Interaction by Child and Condition

	Condition									
	Baseline		Teaching		New recipient		Script fading		New activities	
Type of interaction	Range	M	Range	M	Range	M	Range	M	Range	M
Scripted										
David			88–100	94		100		100		
Jeremiah			89–100	98		100		—[a]		
Ben			80–100	98		90		—[a]		
Elaborations										
David			75–100	95	87–88	88	90–100	95		
Jeremiah			83–100	95	92–100	96		—[a]		
Ben			75–100	95		95		—[a]		
Unscripted										
David		100		100	87–88	88	97–100	98	98–100	99
Jeremiah		100		100	88–100	94	93–100	96		98
Ben		100		100		100	92–93	92	98–100	99

[a]No data are available for scripted responses and elaborations in the script-fading condition, because interobserver agreement was obtained after scripts were removed (i.e., because scripts were absent, there could be only unscripted responses).

From "Social Interaction Skills for Children with Autism: A Script-Fading Procedure for Beginning Readers," by P. J. Krantz and L. E. McClannahan, 1998, *Journal of Applied Behavior Analysis, 31,* p. 196. Copyright 1998 by the Society for the Experimental Analysis of Behavior, Inc. Reprinted by permission.

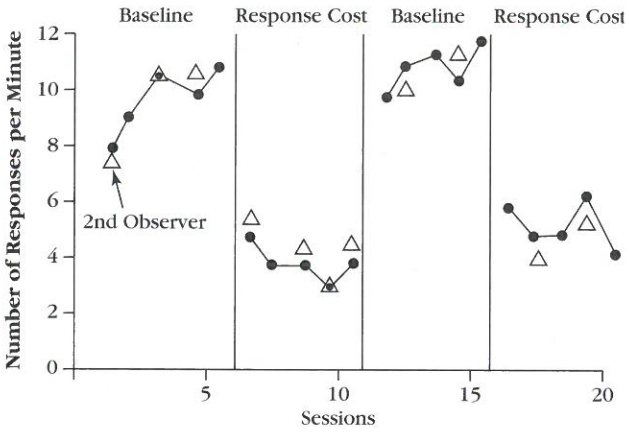

Figure 5.6 Plotting measures obtained by a second observer on a graph of the primary observer's data provide a visual representation of the extent and nature of interobserver agreement.

data in Figure 5.6. When the data reported by the primary observer show clear change in the behavior between conditions or phases and all of the measures reported by the secondary observer within each phase fall within the range of observed values obtained by the primary observer, confidence increases that the data represent actual changes in the behavior measured rather than changes in the primary observer's behavior due to drift or extra-experimental contingencies.

Although published research reports in applied behavior analysis seldom include graphic displays of IOA measures, creating and using such displays during a study is a simple and direct way for researchers to detect patterns in the consistency (or inconsistency) with which observers are measuring behavior that might be not be as evident in comparing a series of percentages.

Which Approach Should Be Used for Assessing the Quality of Measurement: Accuracy, Reliability, or Interobserver Agreement?

Assessments of the accuracy of measurement, the reliability of measurement, and the extent to which different observers obtain the same measures each provide different indications of data quality. Ultimately, the reason for conducting any type of assessment of measurement quality is to obtain quantitative evidence that can be used for the dual purposes of improving measurement during the course of an investigation and judging and convincing others of the trustworthiness of the data.

After ensuring the validity of what they are measuring and how they are measuring it, applied behavior analysts should choose to assess the accuracy of measurement whenever possible rather than reliability or interobserver agreement. If it can be determined that all measurements in a data set meet an acceptable accuracy criterion, questions regarding the reliability of measurement and interobserver agreement are moot. For data confirmed to be accurate, conducting additional assessments of reliability or IOA is unnecessary.

When assessing the accuracy of measurement is not possible because true values are unavailable, an assessment of reliability provides the next best quality indicator. If natural or contrived permanent products can be archived, applied behavior analysts can assess the reliability of measurement, allowing consumers to know that observers have measured behavior consistently from session to session, condition to condition, and phase to phase.

When true values and permanent product archives are unavailable, interobserver agreement provides a level of believability for the data. Although IOA is not a direct indicator of the validity, accuracy, or reliability of measurement, it has proven to be a valuable and useful research tool in applied behavior analysis. Reporting interobserver agreement has been an expected and required component of published research in applied behavior analysis for several decades. In spite of its limitations, "the homely measures of observer agreement so widely used in the field are exactly relevant" (Baer, 1977, p. 119) to efforts to develop a robust technology of behavior change.

> Percentage of agreement, in the interval-recording paradigm, does have a direct and useful meaning: how often do two observers watching one subject, and equipped with the same definitions of behavior, see it occurring or not occurring at the same standard times? The two answers, "They agree about its occurrence X% of the relevant intervals, and about its nonoccurrence Y% of the relevant intervals," are superbly useful. (Baer, 1977, p. 118)

There are no reasons to prevent researchers from using multiple assessment procedures to evaluate the same data set. When time and resources permit, it may even be desirable to include combinations of assessments. Applied behavior analysts can use any possible combination of the assessment (e.g., accuracy plus IOA, reliability plus IOA). In addition, some aspects of the data set could be assessed for accuracy or reliability while other aspects are assessed with IOA. The previous example of accuracy assessment reported by Brown and colleagues (1996) included assessments for accuracy and IOA. Independent observers recorded correct and incorrect student-delayed retells. When IOA was less than 100%, data for that student and session were assessed for accuracy. IOA was used as an assessment to enhance believability, and also as a procedure for selecting data to be assessed for accuracy.

Summary

Indicators of Trustworthy Measurement

1. To be most useful for science, measurement must be valid, accurate, and reliable.
2. Valid measurement in ABA encompasses three equally important elements: (a) measuring directly a socially significant target behavior, (b) measuring a dimension of the target behavior relevant to the question or concern about the behavior, and (c) ensuring that the data are representative of the behavior under conditions and during times most relevant to the reason(s) for measuring it.
3. Measurement is accurate when observed values, the data produced by measuring an event, match the true state, or true values, of the event.
4. Measurement is reliable when it yields the same values across repeated measurement of the same event.

Threats to Measurement Validity

5. Indirect measurement—measuring a behavior different from the behavior of interest—threatens validity because it requires that the researcher or practitioner make inferences about the relationship between the measures obtained and the actual behavior of interest.
6. A researcher who employs indirect measurement must provide evidence that the behavior measured directly reflects, in some reliable and meaningful way, something about the behavior for which the researcher wishes to draw conclusions.
7. Measuring a dimension of the behavior that is ill suited for, or irrelevant to, the reason for measuring the behavior compromises validity.
8. Measurement artifacts are data that give an unwarranted or misleading picture of the behavior because of the way measurement was conducted. Discontinuous measurement, poorly scheduled observations, and insensitive or limiting measurement scales are common causes of measurement artifacts.

Threats to Measurement Accuracy and Reliability

9. Most investigations in applied behavior analysis use human observers to measure behavior, and human error is the biggest threat to the accuracy and reliability of data.
10. Factors that contribute to measurement error include poorly designed measurement systems, inadequate observer training, and expectations about what the data should look like.
11. Observers should receive systematic training and practice with the measurement system and meet predetermined accuracy and reliability criteria before collecting data.
12. Observer drift—unintended changes in the way an observer uses a measurement system over the course of an investigation—can be minimized by booster training sessions and feedback on the accuracy and reliability of measurement.
13. An observer's expectations or knowledge about predicted or desired results can impair the accuracy and reliability of data.
14. Observers should not receive feedback about the extent to which their data confirm or run counter to hypothesized results or treatment goals.
15. Measurement bias caused by observer expectations can be avoided by using naive observers.
16. Observer reactivity is measurement error caused by an observer's awareness that others are evaluating the data he reports.

Assessing the Accuracy and Reliability of Behavioral Measurement

17. Researchers and practitioners who assess the accuracy of their data can (a) determine early in an analysis whether the data are usable for making experimental or treatment decisions, (b) discover and correct measurement errors, (c) detect consistent patterns of measurement error that can lead to the overall improvement or calibration of the measurement system, and (d) communicate to others the relative trustworthiness of the data.
18. Assessing the accuracy of measurement is a straightforward process of calculating the correspondence of each measure, or datum, assessed to its true value.
19. True values for many behaviors of interest to applied behavior analysts are evident and universally accepted or can be established conditionally by local context. True values for some behaviors (e.g., cooperative play) are difficult because the process for determining a true value must be different from the measurement procedures used to obtain the data one wishes to compare to the true value.
20. Assessing the extent to which observers are reliably applying a valid and accurate measurement system provides a useful indicator of the overall trustworthiness of the data.
21. Assessing the reliability of measurement requires a natural or contrived permanent product so the observer can remeasure the same behavioral events.
22. Although high reliability does not confirm high accuracy, discovering a low level of reliability signals that the data are suspect enough to be disregarded until problems in the measurement system can be determined and repaired.

Using Interobserver Agreement to Assess Behavioral Measurement

23. The most commonly used indicator of measurement quality in ABA is interobserver agreement (IOA), the degree

to which two or more independent observers report the same observed values after measuring the same events.

24. Researchers and practitioners use measures of IOA to (a) determine the competence of new observers, (b) detect observer drift, (c) judge whether the definition of the target behavior is clear and the system not too difficult to use, and (d) convince others of the relative believability of the data.

25. Measuring IOA requires that two or more observers (a) use the same observation code and measurement system, (b) observe and measure the same participant(s) and events, and (c) observe and record the behavior independent of influence by other observers.

26. There are numerous techniques for calculating IOA, each of which provides a somewhat different view of the extent and nature of agreement and disagreement between observers.

27. Percentage of agreement between observers is the most common convention for reporting IOA in ABA.

28. IOA for data obtained by event recording can be calculated by comparing (a) the total count recorded by each observer per measurement period, (b) the counts tallied by each observer during each of a series of smaller intervals of time within the measurement period, or (c) each observer's count of 1 or 0 on a trial-by-trial basis.

29. Total count IOA is the simplest and crudest indicator of IOA for event recording data, and exact count-per-interval IOA is the most stringent for most data sets obtained by event recording.

30. IOA for data obtained by timing duration, response latency, or interresponse time (IRT) is calculated in essentially the same ways as for event recording data.

31. Total duration IOA is computed by dividing the shorter of the two durations reported by the observers by the longer duration. Mean duration-per-occurrence IOA is a more conservative and usually more meaningful assessment of IOA for total duration data and should always be calculated for duration-per-occurrence data.

32. Three techniques commonly used to calculate IOA for interval data are interval-by-interval IOA, scored-interval IOA, and unscored-interval IOA.

33. Because it is subject to random or accidental agreement between observers, interval-by-interval IOA is likely to overestimate the degree of agreement between observers measuring behaviors that occur at very low or very high rates.

34. Scored-interval IOA is recommended for behaviors that occur at relatively low frequencies; unscored-interval IOA is recommended for behaviors that occur at relatively high frequencies.

35. IOA assessments should occur during each condition and phase of a study and be distributed across days of the week, times of day, settings, and observers.

36. Researchers should obtain and report IOA at the same levels at which they report and discuss the results of their study.

37. More stringent and conservative IOA methods should be used over methods that may overestimate agreement as a result of chance.

38. The convention for acceptable IOA has been a minimum of 80%, but there can be no set criterion. The nature of the behavior being measured and the degree of behavior change revealed by the data must be considered when determining an acceptable level of IOA.

39. IOA scores can be reported in narrative, table, and graphic form.

40. Researchers can use multiple indices to assess the quality of their data (e.g., accuracy plus IOA, reliability plus IOA).