Behavior Analysis: Research and Practice
2016, Vol. 16, No. 1, 19–33

2372-9414/16/$12.00 http://dx.doi.org/10.1037/bar0000031

Data Collection and Measurement Assessment in Behavioral Research: 1958–2013

Douglas E. Kostewicz
University of Pittsburgh

Seth A. King
Tennessee Technological University

Shawn M. Datchuk
University of Iowa

Kaitlyn M. Brennan
University of Pittsburgh

Sean D. Casey
Iowa Department of Education, Des Moines, Iowa

The measurement of behavior plays an integral role in psychology and its subfields such as behavior analysis. Behavior analysts, as with all scientists, must establish a clear and concise link between observed measures and the actual phenomena under observation. Three measures help establish the link—interobserver agreement, reliability, and accuracy. Authors in the current review surveyed over 2,000 studies from behavioral journals published between 1958 and 2013. Guiding questions covered how behavior analysts collect data and to what extent and how do they conduct assessments of the dependent variables. Results indicated that the collection of data across behavior analytic research occurs equitably between direct observation, permanent product, and automated recording. In addition, only a third of studies include dependent measure assessment with the vast majority occurring at the interobserver agreement level. The discussion centers on issues surrounding the reliance on interobserver agreement within our science and the potential of future technological advancements to improve the link between measurement and the natural world.

Keywords: accuracy, reliability, interobserver agreement, data collection

A fundamental quality of science involves the precise and meaningful measurement of observed phenomena (Skinner, 1953). Observations serve as the backbone to form or test scientific theories, perpetuate new discoveries, and increase the understanding of the natural world. Skinner (1939) displayed over 75 years ago that the precise measurement of behavior led to core behavior analytic principals such as reinforcement and punishment. Such assiduous measurement distinguishes the analysis of behavior and other forms of scientific inquiry from less rigorous forms of observation (Johnston & Pennypacker, 2009; Matthews, 1998; Skinner, 1953). Fundamental standards of quality (e.g., Horner et al., 2005; Hudson, Lewis, Stichter, & Johnson, 2011) require behavior researchers to present operationally explicit and technologically replicable measurement systems (Ayres & Gast, 2010; Baer, Wolf, & Risley, 1968).

Measurement assessment, or the evaluation of error in scientific data, represents an additional concern for behavioral scientists due to measurement uncertainty (Joint Committee for Guides in Metrology [JCGM], 2008) stemming from factors (e.g., observer error, time sampling error) that obscure the relationship between values obtained through measurement and true values—the actual dimensions of the phenomenon

Douglas E. Kostewicz, Department of Instruction and Learning, University of Pittsburgh; Seth A. King, Department of Curriculum and Instruction, Tennessee Technological University; Shawn M. Datchuk, Department of Teaching and Learning, University of Iowa; Kaitlyn M. Brennan, Department of Instruction and Learning, University of Pittsburgh; Sean D. Casey, Iowa Department of Education, Des Moines, Iowa.

Correspondence concerning this article should be addressed to Douglas E. Kostewicz, Department of Instruction and Learning, University of Pittsburgh, 5162 Posvar Hall, Pittsburgh, PA 15260. E-mail: dekost@pitt.edu

of interest (Johnston & Pennypacker, 2009; Kazdin, 1977; Kazdin, 2012; Rapp, Carroll, Stangeland, Swanson, & Higgins, 2011). In terms of measurement, "true value" represents a standard value or referent generally regarded as less susceptible to the error encumbering all measurement systems, rather than an unattainable theoretical concept (i.e., value determined in the absence of error; Buekens & Truyen, 2014). Values derived from measurement inexorably differ from the true value due to systematic error, or bias (e.g., a speedometer that consistently provides a true speed of 60 mph as 55 mph), and random errors, or unpredictable influences on the measurement procedure (e.g., the influence of temperature changes on the speedometer; Farrance & Frenkel, 2012). Thus, researchers consign themselves to accounting for potential sources of error in their measures, often using indicators of quality measurement including: accuracy and reliability (Cooper, Heron, & Heward, 2007).

Accuracy, the disparity in values observed by researchers and a true value (e.g., standard or referent measure) of the object of interest, represents the highest form of measurement assessment (Cooper et al., 2007). Procedures in laboratory and other mature sciences generally entail calculating a range of values based on repeated measures and probabilistic statistical models (see JCGM, 2008; White, Farrance, & the AACB Uncertainty of Measurement Working Group, 2004). Guidelines for the determination of accuracy in clinical disciplines often compare observed values to measures obtained through procedures that (a) differ from the initial observation methods and (b) eliminate potential sources of error (Johnston & Pennypacker, 2009; White et al., 2004). Examples of such procedures include comparing the self-monitored blood glucose estimates reported by patients with diabetes to levels determined through a glucose analyzer (Clarke, Cox, Gonder-Frederick, Carter, & Pohl, 1987). Highly accurate data may serve as the basis for decisions in research and in practice (Johnston & Pennypacker, 2009; Wolery, 2004).

Imprecise data merits consideration, however, provided the observed values fall within a reasonable range of values and reflect a relatively consistent pattern of error (Johnston & Pennypacker, 2009). *Reliability* refers to the extent to which repeated measurements of the same event produce identical values (Cooper et al., 2007). Johnston and Pennypacker (2009) suggested that researchers derive reliability by repeatedly observing a static sample of the phenomenon of interest. For example, Andersen, Jorgenson, and Hededam (1990) evaluated the use of a diagnostic system for hip fractures by requiring six participants (i.e., consultants and surgeons in training) to independently evaluate 49 radiographs depicting various fractures. The researchers obtained a reliability coefficient by requiring participants to review the same set of radiographs at two separate times. No less than 1 month of time separated the initial and follow-up review sessions, yielding an average Kappa statistic of 0.74 (range = 0.66–0.92) Although not sufficient to infer accuracy, reports of reliability provide additional evidence related to the calibration of measurement (Johnston & Pennypacker, 2009; Wu, Whiteside, & Neighbors, 2007).

Assessing accuracy and reliability contributes to the quality of evidence presented through research (Cooper et al., 2007). Obtaining a comparative referent to establish accuracy or reliability poses a potential challenge to researchers and practitioners in the behavioral fields, however, given the historical prevalence of direct observation in behavior science (Boykin & Nelson, 1981; Kelly, 1977; Mitchell, 1979). Consequently, researchers generally assess measures using interobserver agreement (IOA), the comparison of two independent observations of a single event (Kazdin, 1977; Watkins & Pacheco, 2000). Although synonymous with measurement assessment in much of the behavioral literature, the extent to which IOA represents a viable indicator of accuracy or reliability remains a point of contention (Brington, Palmer, & Watkins, 2002; Crowley-Koch & Van Houten, 2013; Mudford, Zeleny, Fisher, Klum, & Owen, 2011).

IOA: Issues and Concerns

Kazdin (1977) observed that the use of IOA as the primary form of measurement assessment in behavioral research often occurs under the assumption that accuracy may be inferred from agreement. Nonetheless, researchers generally dismiss the connection between accuracy and IOA entirely (e.g., Kennedy, 2005) or, at best, acknowledge a tenuous link between the two

concepts (Boykin & Nelson, 1981; Deitz, 1988). In delineating the various terms used to identify concepts related to measurement assessment in behavioral studies, Suen (1988) established clear distinctions in the conceptual notions of accuracy obtained through alternate "incontrovertible" (p. 358) assessment of the target phenomenon and reliability as derived through IOA. Johnston and Pennypacker (2009) adamantly rejected IOA as an indicator of accuracy given the lack of evidence to support the accuracy of additional observers:

> The fact that two observers reported the same measure of the target behavior for a session . . . [and] does not provide any guarantee that either report reflects exactly what happened with the target behavior. Such information merely encourages the investigator and others to accept the data as believable. (p. 149)

The susceptibility of IOA to bias compounds the conceptual issues related to accuracy (Fradenburg, Harrison, & Baer, 1995). Research has repeatedly demonstrated the reactivity of IOA to a variety of factors including the quality of observer training (e.g., Boykin & Nelson, 1981), the complexity of behavior definitions (e.g., Dorsey, Nelson, & Hayes, 1986), and covert decline (i.e., changes in IOA when secondary observations occur unbeknownst to the primary observer; Weinrott & Jones, 1984). In addition, IOA varies considerably as a function of the algorithm experimenters use to compare secondary and primary observations (e.g., Watkins & Pacheco, 2000).

Despite consensus regarding the inability of IOA to inform assessments of accuracy and potential sources of bias, various interpretations of the value of assessing reliability through IOA exist. Johnston and Pennypacker (2009) deemed IOA as insufficient means of assessing reliability and identified *intraobserver agreement*—wherein a single observer repeatedly scores a video recording or sample of a single target behavior—as the preferred evaluation method. Following an experimental comparison of IOA and intraobserver agreement, Boyce, Carter, and Neboschick (2000) cautiously concluded that (a) the inclusion of multiple observers to observation sessions (i.e., conducting IOA) contributed little value to the assessment of reliability and (b) the advent of accessible video recording technology could potentially contribute to an increase in reports of intraobserver agreement. In contrast, Suen (1988) identified IOA as a method of determining reliability and favorably compared the procedure with intraobserver agreement, noting that, in practice, intraobserver agreement does not accurately recreate the conditions of the original measurement context. Suen and others (e.g., Cone, 1977; Mitchell, 1979) also noted that conducting both IOA and intraobserver agreement can assist researchers in identifying the sources of error in observation data.

Measurement quality directly relates to the ability of behavior scientists to make credible claims regarding the efficacy of treatment (McDermott, 1988). IOA represents an imperfect approach to measurement assessment that, though unrelated to accuracy and highly subject to bias, remains vital to the continued use of direct observation (Watkins & Pacheco, 2000). The integrity of behavior research findings directly relates to the extent to which reports of measurement assessment appear in the research. As noted by Johnston and Pennypacker (2009), IOA merely contributes to the "believability" of findings reported in behavior studies, and though technically a form of measurement assessment, does not compare to the measures of accuracy featured in more mature sciences. Approaches to measurement assessment may adapt, however, based on growing concerns with methodology and the emergence of technology to facilitate more accurate measurement (e.g., Mudford, Taylor, & Martin, 2009).

As illustrated above, the assessment of dependent measures hardly represents a novel area of concern among researchers (Kelly, 1977; Neely, Davis, Davis, & Rispoli, 2015). Likewise, experimenters have long-held misgivings regarding the practice of relying on IOA as the sole assessment of dependent measures (Crowley-Koch & Van Houten, 2013; Mudford et al., 2011; Kazdin, 1977). Nonetheless, a well known "incongruity between need and action" often occurs within science (Makel & Plucker, 2014, p. 306). The extent to which behavior researchers address calls to assess measurement and pursue alternatives to IOA remains unclear.

The problem distills the relationship between how behavior analytic researchers collect data and in turn assess the data retrieved. Changes in the technology of data collection (e.g., compact video recorders) may lead to more precise assessment of dependent measures, practices within the field—and, therefore, more defensible results. Thus, the current study reviews data

collection and measurement assessment methods in the field of behavior research. Specific questions include (1) To what extent do behavior researchers across subdisciplines (e.g., experimental, applied, and behavior therapy) collect data using direct observation, evaluation of permanent products, and/or automated recording; in addition, how often do behavior researchers employ audio/visual (A/V) recording methods in collecting data? (2) How often do behavior researchers across subdisciplines report assessing measurements of the DV and to what extent has the reports of measurement assessment data changed over time? (3) To what extent do behavior researchers across subdisciplines evaluate dependent measures via IOA, reliability, and/or accuracy and has that changed over time? In addition, how and to what extent do behavior researchers assess dependent measures across various data collection methods.

Method

The journal selection and scoring process included three distinct phases. The first phase involved the identification of behavior journals, placement of journals into categories, and the selection of issues for review. Authors then developed a series of indicators designed to evaluate the collection and measurement assessment of the dependent variables (DV) featured within the research sample. For the final phase, scorers assessed selected issues of behavioral research using the evaluation indicators.

Journal, Issue, and Article Selection

Previous literature analyses (Carr & Britton, 2003; Critchfield, 2002; Kubina, Kostewicz, & Datchuk, 2008) provided the initial journal identification criteria. The search identified six journals that examine technical applications, practices, and issues related to the field of behavior analysis (see Table 1). To increase the representative nature of the review, the final sample included an additional six journals with a 10-year publication record that pertained to the application of behavior-analytic principles with additional emphases on education, experimental behavior analysis, and the analysis of verbal behavior (see Table 1).

A team of three doctoral-level behavioral researchers divided the journals into three categories corresponding with the primary dimensions of behavior analysis: experimental, applied, (Cooper et al., 2007) and behavior therapy (i.e., eclectic, cognitive–behavioral; Carr & Britton, 2003; Critchfield, 2002). The reviewers determined category designation based on the stated mission and content of each journal. Journal category and placement appear on Table 1.

Authors then randomly selected one issue from every two consecutive volumes per journal title published from inception to 2013. Within each identified issue, article inclusion criteria encompassed experimental or case studies that presented previously unpublished data from human or animal participants and maintained a clearly defined methods section. Non-

Table 1
Journal Titles and Categories

Category	Journal title
Experimental	*Journal of Experimental Analysis of Behavior*[b]
	Learning and Behavior[b]
Applied	*Journal of Applied Behavior Analysis*[a]
	Behavior Modification[a]
	Education and Treatment of Children[b]
	European Journal of Behavior Analysis[b]
	Analysis of Verbal Behavior[b]
	Journal of Behavioral Education[b]
Behavior therapy	*Cognitive Behavioral Practice*[a]
	Behavior Therapy[a]
	Journal of Behavior Therapy and Experimental Psychiatry[a]
	Child and Family Behavior Therapy[a]

[a] Journal identified during initial search. [b] Journal identified to gather a representative sample.

research articles including editorials, literature reviews, commentaries, letters to the editor, book reviews, obituaries, news updates, and reviews of products or pharmaceuticals met exclusion criteria (Kubina et al., 2008). For included articles containing one or more studies, each study received distinct scores based on the description of methods provided for each. Exceptions involved (a) the provision of a general methods section or (b) the explicit reported use of identical methods across studies.

DV Indicators

An expert panel consisting of three PhD faculty members with over 50 years combined experience in behavioral research design developed the scoring indicators applied to the collection and assessment of DVs within each identified study. Dependent measures or DVs refer to data collected that attempt to address the stated research questions. Exceptions eliminated from review included any additional measures (e.g., implementation fidelity, preassessment or screening surveys, and social validity). The completed DV indicators encompassed three subcategories shown on Table 2: data collection method, data collection assessment, and data collection assessment method.

Method of data collection. Data collection indicators denoted the presence or absence of techniques used to collect DVs as well as instances in which the experimenters collected data through automated recording, evaluation of permanent products, or direct observation (see Table 2). Automated recording methods included mechanical devices, such as scales, di-

Table 2
Coding Indicators Across Data Collection Method, Assessment, and Assessment Method

Indicator	Subcategories	Description
Data collection method	Automated	Any mechanical device used to measure a dependent variable (DV).
	Permanent product	Events detected from changes in the environment as indirect representations of the DV. Included instances of audio/visual recording to preserve evidence of the DV.
	Direct observation	Any data collection technique in which the researchers collected data through in-situ observation of the DV.
Data collection assessment	Complete assessment, full report	Study reported (a) an assessment of all DVs and (b) provided a process for assessing the DV, resulting score and formula, and the percent of sessions assessed.
	Complete assessment, partial report	Study reported (a) an assessment of all DVs and (b) withheld some or all aspects noted in a full report.
	Incomplete assessment, full or partial report	Study reported (a) an assessment of some of the DVs and (b) provided either a full or partial description of the assessment process.
	No reported assessment	Study did not report an assessment of the DV.
Data collection assessment method	Accuracy	Comparisons of the score from an observation to (a) the results of a more rigorous scoring procedure unrelated to the experimental context and/or (b) the 'true value' of the observed phenomena (Johnston & Pennypacker, 2009)
	Reliability	Assessment of the consistency of measurement over multiple replications through either (a) the repeated presentation of the same sample of the DV to a single observer or (b) the use of assessment techniques designed to determine the consistency of a test/instrument.
	Interobserver agreement	Comparison of the independent observations of the same event by two or more observers.

rectly involved in obtaining quantities of interest with minimal observer maintenance (e.g., shifting scale balance weights). However, automated measurement excluded researcher-operated devices used to aggregate data collected by observers (e.g., Multi-Option Observation System for Experimental Studies; Tapp, 1995). Permanent products included the observation of events detected through changes in environment or other artifacts produced following the occurrence of a behavior (e.g., evidence of self-injurious behavior, surveys). Direct observation included data collected through in situ observation of a phenomenon and self-report of observable phenomena as well as tests involving responses recorded by the experimenters (e.g., eye exams) and other forms of recording data (e.g., height measurement) for which the authors did not specify the measurement technique. Although potentially susceptible to transcription errors, direct observations did not include instances in which the DV reflected observations of a permanent product (e.g., counting school yard graffiti) or observations of data generated by automated recording devices (e.g., observation of a digital scale).

Data collection assessment. Data collection assessment indicators (see Table 2) pertained to the reported assessment of all featured dependent measures and the level of detail provided. Notwithstanding method specific exceptions (see below), full reports included all of the following: a score related to the quality of measurement (e.g., percent IOA), description of process and formulas used to obtain the rating score, and a percentage of sessions or samples in which researchers assessed measurement. Partial reports either (a) lacked some of the information required to qualify for the full report designation or (b) provided no evidence of measurement assessment beyond a nominal reference to a specific data collection assessment method.

Data collection assessment method. Data collection assessment method indicators (see Table 2) differentiated the methods used by experimenters to assess DVs and determine the level of reporting (i.e., partial or full). The indicators corresponded with the three broad methods for assessing measurement: IOA, reliability, and accuracy. When experimenters reported conducting multiple forms of measurement assessment, the study received scores for each assessment method. When possible, assessment methods received scores based on reported procedures rather than the label provided within the text.

Data collection through the use of mechanical instruments did not receive credit for accuracy in the absence of corroborative measures (e.g., calibration certificate). However, experimenters who described calibration procedures received credit for fully reporting accuracy. Studies in which the experimenters assigned a "master" observer (e.g., lead researcher) to rescore video recordings of session data received credit for reporting IOA.

Given the frequent use of psychometric instruments in behavior therapy settings (e.g., Sturmey, 1994), measures of test/instrument assessment techniques qualified as evidence of reliability, provided the experimenter conducted all procedures within the context of the study. The reporting of previously published instrument assessments did not meet the current standards of reliability. Studies featuring the use of instrument assessment received credit for fully reporting reliability provided the authors specifically identified the procedure.

Assessment of Selected Research

Prior to coding, three reviewers scored behavior research studies until agreement, defined as all reviewers recording the same entry for a single category in each study, exceeded 90%. Studies used during the training process originated from journals outside of the sample. Reviewers continued to score different studies until meeting criterion. Thereafter, the reviewers applied the assessment indicators to individual studies appearing in articles from the selected issues. Coding occurred on Excel spreadsheets with each row representing an experiment. Coders read the dependent measure and measurement assessment section of each article and scored each column as appropriate. Codes encompassed correspondence between all dependent measures and assessment procedures described in a given study. Because of space limitations of many publications, scoring of additional material occurred when experimenters cited supplemental descriptions featured in another peer-reviewed study.

IOA agreement. To determine IOA, a second reviewer assessed 21% of reviewed issues (i.e., all studies meeting criteria within the issue). The second author calculated IOA using

an exact agreement (i.e., point-by-point) approach (Kennedy, 2005), or dividing the indicators in which observers recording the same value by the total number of indicators per scored study and multiplying by 100. IOA across studies averaged 90% (range = 76–100%). The second author resolved all interobserver discrepancies.

Reliability. Reviewers without forewarning completed a reassessment of 20% of previously scored issues throughout the research review process. Approximately 30 days elapsed between scoring and rescoring. The second author calculated reliability by dividing the percentage of indicators in which the reviewers' initial score matched the reassessment score by the total number of indicators and multiplying by 100. Reliability for reassessed issues averaged 97%. The three reviewers achieved average reliability scores of 99% (range = 94–100%), 96% (range = 91–100%), and 91% (range = 75–100%). Individual reviewers resolved intraobserver scoring discrepancies.

Results

The selection process randomly identified 211 issues within the 12 selected journals with an average of 18 issues per journal (range 7–28) published between 1958 and 2013. A total number of 1,516 individual articles met criteria and contained 2,091 individual studies. Averages of 37 (range 8–91) studies met criteria each year of analysis and 174 (range 23–500) per journal title.

Data Collection Method

Figure 1 shows a dot chart representing the number of studies (*x*-axis) that account for the various data collection methods (*y*-axis). Each line contains a breakdown of the three journal groupings (e.g., experimental, applied, and behavior therapy). Distance from the left horizontal axis to the closed dot represents number of studies from experimental (EAB) journals. From the closed dot to the open square, applied (ABA) journals' studies, and from the open square to open circles, behavior therapy (BT) journals' studies.

Authors reported a fairly even breakdown of data collection methods across the three categories. Isolated use of automated (713), direct observation (540), and permanent product (472), comprised 83% of the study sample. The remaining 17% of studies reported mixed methods combining direct observation and either permanent product (200) or automated recording (26) or using automated and permanent product together (111). The final 29 studies contained all three types. Approximately 10% (213) of the study sample presented A/V evidence of behaviors across all data collection methods.

Analyzing results by journal category reveals key differences. EAB, ABA, and BT journals emphasized one of the three primary data collection methods. Automated recording methods appeared in 69% (624) of studies in the EAB category. Researchers in 47% (329) of ABA journals' studies reported using direct observation in isolation.

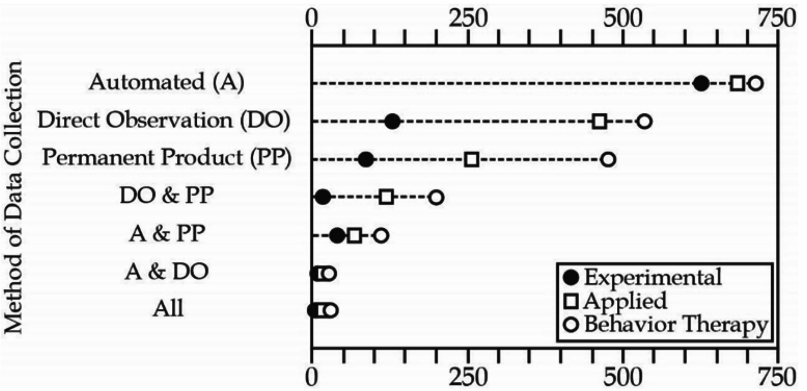

Figure 1. Reported method(s) of data collection within identified studies across journal type.

Out of 488 studies in BT journals, 45% used permanent product.

Data Collection Assessment

Overall, procedures containing either a full (221) or partial (245) report for measurement assessment comprised 22% of studies. A smaller percentage (11%) of method sections (249) included an incomplete (i.e., did not report assessing all measures) assessment. The remaining 66% of the total studies (1,376) did not report conducting any form of measurement assessment.

The line chart in Figure 2 illustrates the instances of reporting (*y*-axis) that has occurred in continuous four year blocks of time from 1958 up to and including 2013 (*x*-axis). Solid circles represent instances of no reported assessment and solid squares show complete, full reports. Open squares and circles display partially complete and incomplete reports, respectively.

After an initial increase, instances of no reported assessment decreased steadily from 1974 (149) until 2005 (81) culminating in 118 after an increase over the last eight years. Cases of the three remaining categories appeared more often until 1978 remaining relatively stable until 1997. Complete full reports then increased by a factor of nine (7 to 67) over the sample's final 14 years with complete partial reports decreasing by a factor of seven (37 to 5). Incomplete reports remained stable over the sample's final years following a doubling (18 to 37) in 1998.

Figure 3 contains three line charts focusing on measurement assessment in each of the journal categories. Each of the line graphs within Figure 3 follow the conventions established on Figure 2, except open circles represent the combination of full, partial, and incomplete reports within each journal category. Studies in EAB journals without dependent measure assessment peaked in 1981 with 151 and have declined steadily since. Over the 55 year sample, instances of EAB dependent measure assessment remained stable at approximately two peaking at nine in 1998–2001. ABA journals increased in both reported assessment and no assessment of DVs over the past 44 years with reports of DV proportionally higher than no assessment every four year block. After an initial increase, studies in BT journals containing no assessment of DV slowly declined. Conversely, instances of DV assessment increased until 1998. At that point, reports of assessment remained stable with no reports of assessment increasing.

Data Collection Assessment Method

A total of 715 studies out of 2,091 (34%) described employing some form of data collection assessment procedures. Occurrences of

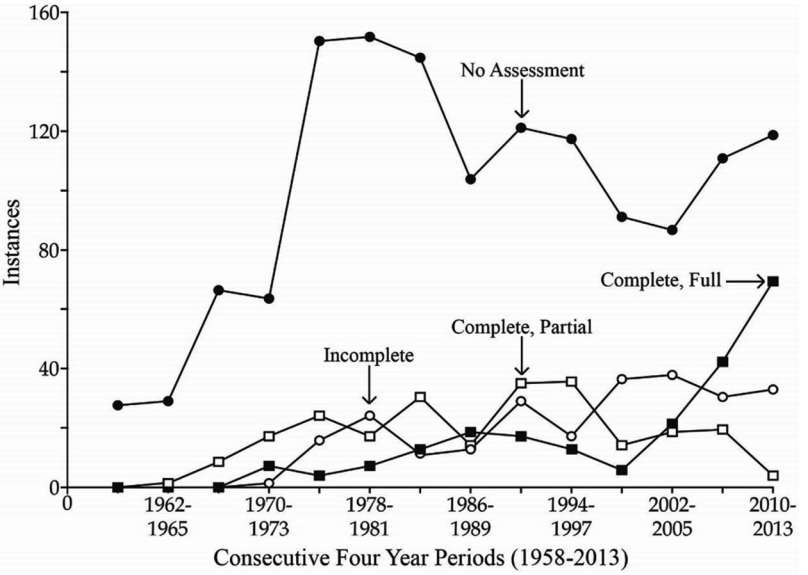

Figure 2. Dependent measure assessment reporting from 1958 to 2013.

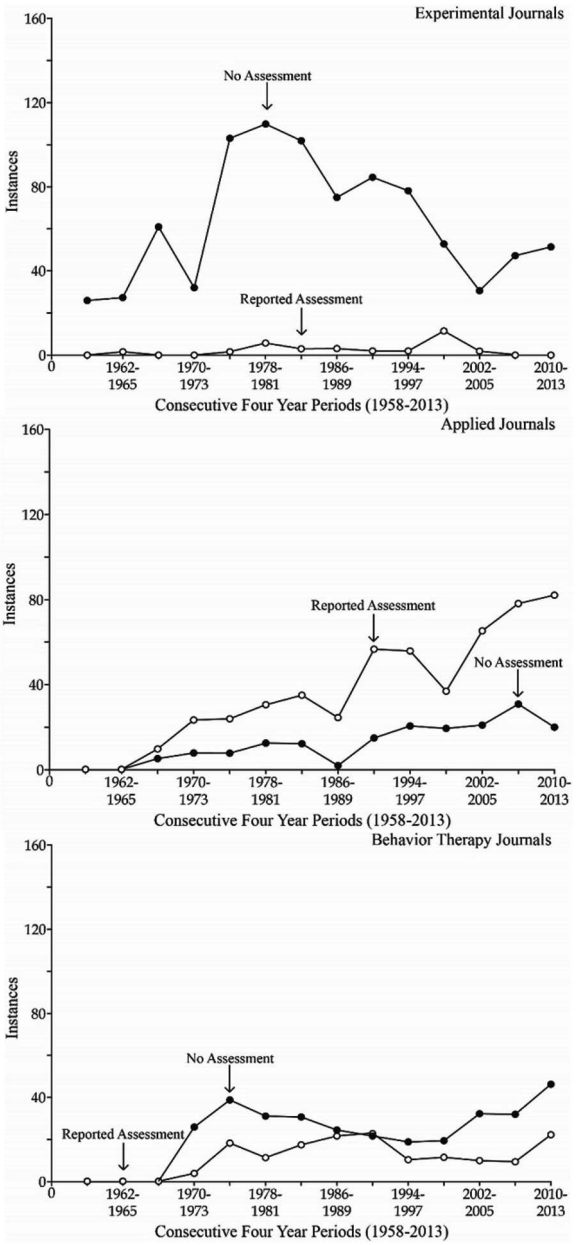

Figure 3. Dependent measure assessment reporting by journal type from 1958 to 2013.

IOA (628) comprised the bulk of measurement assessment and made up 30% of the total sample. Accuracy (16) and reliability (40) accounted for the sole DV assessment measure in a small proportion of total studies (3%). Assessment approaches that involved more than one type of procedure occurred rarely (31). The remaining 66% (1,376) of studies contained no reported assessment.

The line chart in Figure 4 shows the various types of data collection assessment methods in continuous four year blocks spanning 1958

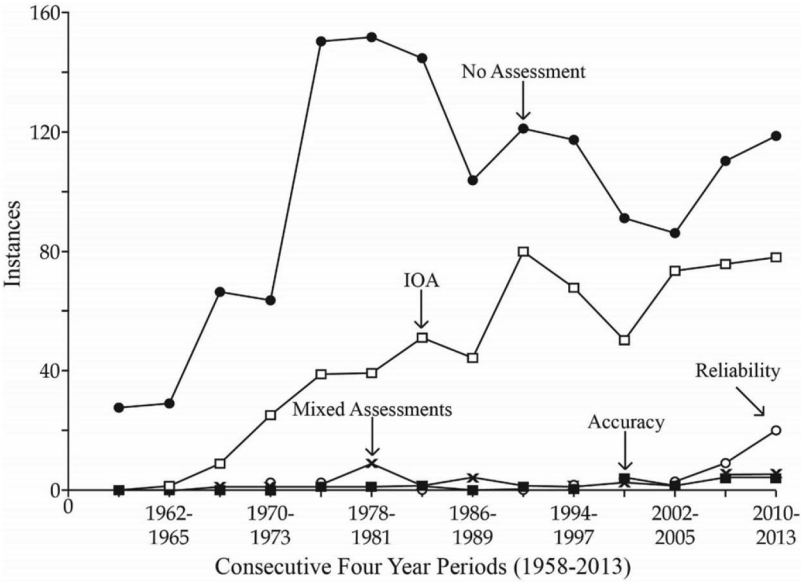

Figure 4. Measure assessment techniques reported from 1958 to 2013.

through 2013. Closed circles, again, represent no reported assessment. Open squares and open circles refer to reports that include only IOA and reliability. Closed squares refer to instances of accuracy and X's mean experimenters reported making use of more than one type of assessment.

No reported assessment occurs on Figure 4 as a reference and follows the same path as noted in Figure 2. Over the same time span, experimenters reported IOA with a steady increase in frequency from 0 in 1958 to almost 74 in 2005. The stabilization of IOA over the remaining eight years coincided with a no assessment resurgence. The remaining three measures (i.e., accuracy, reliability, mixed measures) occurred at rates near 0 for the entire 56 year sample.

Table 3 shows a breakdown of DV assessment type per journal category. ABA authors include a DV assessment in approximately 75% of studies. IOA dwarfs all other types of DV assessment in ABA journals both within and between journal categories. Experimenters in BT journals report assessing DVs in slightly more than a third of instances with IOA (112) and reliability (37) accounting for the majority. EAB journals contain only 3% instances of DV assessment.

Table 4 displays a description of data collection type per DV assessment method. Experimenters report IOA in over 50% of instances when collecting data via direct observation in isolation or conjunction with other types and regardless of journal category. IOA also dominate studies featuring data collecting via permanent product. Authors generally did not rely on any form of DV assessment when collecting data via automated recording.

Table 3
Dependent Measure Assessment Per Journal Type

Journal category	Studies	No assessment (%)	IOA (%)	Reliability (%)	Accuracy (%)	Mixed (%)
Experimental journals	907	883 (97%)	18 (2%)	0 (0%)	5 (.5%)	1 (.1%)
Applied journals	696	174 (25%)	498 (72%)	3 (.4%)	9 (.9%)	12 (2%)
Behavior therapy journals	488	319 (65%)	112 (23%)	37 (8%)	2 (.4%)	18 (4%)
Total	2,091	1,376 (66%)	628 (30%)	40 (2%)	16 (.8%)	31 (1%)

Note. IOA = interobserver agreement.

Table 4
Dependent Measure Assessment Per Data Collection Method

Method of data collection	Studies	No assessment (%)	IOA (%)	Reliability (%)	Accuracy (%)	Mixed (%)
Automated (A)	713	696 (98%)	7 (1%)	0 (0%)	9 (1%)	1 (.1%)
Direct observation (DO)	540	195 (36%)	343 (63%)	0 (0%)	0 (0%)	2 (1%)
Permanent product (PP)	472	299 (63%)	143 (30%)	19 (4%)	1 (.2%)	10 (2%)
DO & PP	200	63 (32%)	108 (54%)	12 (6%)	0 (0%)	17 (9%)
A & PP	111	94 (85%)	7 (6%)	9 (8%)	1 (1%)	0 (0%)
A & DO	26	14 (54%)	9 (35%)	0 (0%)	3 (12%)	0 (0%)
All	29	15 (52%)	11 (38%)	0 (0%)	2 (7%)	1 (3%)
Total	2,091	1,376 (67%)	628 (30%)	40 (2%)	16 (.8%)	31 (1%)

Note. IOA = interobserver agreement.

Summary of Results

Findings revealed a generally equitable distribution of how (i.e., direct observation, permanent product, automated recording) behavior analysts collected data. A small sample of authors also reported collecting A/V evidence of behavior. Reports of measurement assessment did not appear in the bulk of sampled research (66%). The remaining studies featured full or partial descriptions of measurement assessment across all measures (22%) or reported assessing less than the total number of study measures (11%). Examples of measurement assessment techniques increased gradually over time, but the trend of studies that do not report measurement has not abated. When assessed and reported, behavior analysts overwhelming relied on IOA coefficients.

Discussion

Observation and measurement of an organism's interaction with and movement in an environment foundationally serve a science devoted to behavior (Johnston & Pennypacker, 2009). The advancement of behavior analysis, as with all sciences, depends heavily on establishing a strong relationship between reported measurement and actual events (Mudford et al., 2011). Current behavior analysts employ IOA, reliability, and/or accuracy scores as the link between measurement and the observed world (Cooper et al., 2007). The three measures, however, do not share theoretical nor operational equivalence (Suen, 1988). Thus, the current review evaluated the data collection procedures, reports of measurement assessment, and the use of IOA, reliability, and accuracy in a sample (n = 2,091) of behavioral studies published from 1958 to 2013.

Behavior analysts have a variety of tools and techniques available for collecting behavioral data. The methods range from micro switches to behavioral coding sheets. Researchers in the field balance data collection across direct observation, permanent product, and automated recording. Behavioral researchers from different subfields tend to rely on different measurement procedures (see Figure 1). The various data collection methods facilitate certain forms of measurement assessment. Use of direct observation alone, for example, limits approaches to assessing the dependent measure and has reportedly contributed to the prominence of IOA within the behavioral sciences (Watkins & Pacheco, 2000).

Although ostensibly resistant to much of the error inherent in assessing the DV (e.g., observer drift; Dixon, Nastally, Jackson, & Habib, 2009)—the use of automated measurement does not preclude the need for assessing the DV, as authors must ultimately accompany measures with reports related to the veracity of outcomes. Only 17 (2%) of 713 studies using automated recording presented any dependent measure assessment. Skinner (1939) "easily" (p. 311) used known weights and a stopwatch to determine response intensity and duration, respectively. The calibration and accuracy of responding occurred by comparing the graphs obtained through observed behavior (i.e., lever presses) to the graphs obtained from different weights pressed on the lever and the lever held for different set times.

Skinner displayed the important connection between data collection and the assessment of

measures. Automated and permanent product recording, unlike direct observation, facilitates the derivation of reliability and accuracy. Unprecedented access to video recording technology, though observed in only 11% of the identified literature, permits current researchers to conduct accuracy and reliability (i.e., intraobserver agreement). Thus, the spectrum of data collection techniques within the behavior literature and the availability of technology potentially extend the applicability of alternatives to IOA and, in general, data collection assessment methods.

Regardless of collection methods, the presentation of trustworthy data represents a paramount concern within behavior analysis (Baer et al., 1968; Neely et al., 2015; Skinner, 1939). However, two thirds of reviewed studies did not provide data related to the assessment of the DV. Reporting practices over time and between subdisciplines suggest another story. The presence of DV assessment steadily increased since 1958 (see Figure 2). Studies within the ABA subdiscipline account for the majority of the increase (see Figure 3). Two related reasons may better explain the observed trend.

First, editors increasingly specified guidelines for reporting measurement assessment within submitted manuscripts. The *Journal of Applied Behavior Analysis* (*JABA*), as an example and an ABA journal, currently lists IOA as a necessary measure. Earlier editions, however, featured no such guidelines. Second, a series of *JABA* articles in the late 1970s (e.g., Kazdin, 1977; Kelly, 1977) underscored the importance conducting and reporting assessments of dependent measures. ABA researchers appear to have responded by increasingly including DV assessment in published research. Notwithstanding the calls for DV assessment in formative behavioral science texts (Skinner, 1939), the emphasis on DV assessment observed within ABA does not appear within other subdisciplines of behavior analysis (see Figure 3).

Several caveats mitigate the reported increases to measurement assessment. A closer examination reveals that across behavior analysis during any 4-year period authors failed to report assessing the DV in fewer than half of evaluated studies. Of the studies that reported data collection assessment, most either did not thoroughly describe the assessment procedure or did not report assessing all DVs. The rising trend of studies providing a thorough description of all DVs, though heartening, emerged only within the last decade. Kazdin (1977) noted that behavior analysts harbor a "well known concern for consistency and accuracy" in measurement (p. 141). The qualifications notwithstanding recent increases in reported measurement assessment fall considerably short of the presented ideal.

Dependent measure assessment ultimately serves as the quantifiable index of the relation between measures and the phenomena of interest (Johnston & Pennypacker, 2009). As mentioned previously, IOA and accuracy represent two ends of the DV assessment spectrum. IOA serves as a measure comparing the scores of two or more observers. Accuracy measures, on the other hand, represent a concrete relationship between the observed score and the value of measured quantity (Johnston & Pennypacker, 2009). The data on DV assessment present in the behavioral literature appears counter intuitive to notions of data collection consistency and accuracy (Kazdin, 1977). DV assessment occurs exclusively at the IOA level across behavioral subdisciplines (see Table 3) and data collection techniques (see Table 4).

The absence of permanent representations of behavior through direct observation precludes the calculation of reliability and accuracy (Johnston & Pennypacker, 2009). An equitable distribution of data collection techniques (e.g., permanent product, A/V, automated recording) observed in the evaluated studies, however, allow for more informative dependent measure assessment. Behavior analysts have an increasing ability to derive reliability and accuracy given the data collection variety and technological advances. Regardless, various iterations of IOA remain the popular, but questionable, choice when assessing dependent measures.

Epistemologically speaking, IOA assesses the behavior of observers rather than the phenomena under observation, and therefore represents a less robust means of affirming observations of external events. It does not follow that studies incorporating IOA necessarily produce false conclusions, or that—when resources prevent a more thorough examination of the DV—efforts should not be made to improve the calculation and collection IOA (e.g., discontinuing the use of total agreement, use of blind observers). Nonetheless, two simultaneous indepen-

dent observations of a phenomenon, though certainly worthy of more consideration than an individual sighting, ultimately does not provide the most rigorous assessment of the dependent measure.

Limitations

Because of the sampling method applied in the current study, the results do not represent the totality of research published within the behavior sciences. The initial literature search did not include journals involving autism spectrum disorders and other developmental disabilities that frequently publish behavior-based, single-case research designs. Therefore, though the methods of the current study remain consistent with previously published literature surveys, these search limitations must be taken into consideration when interpreting the results.

The three broad categories of behavior analysis employed in the current study (i.e., experimental, applied, and behavior therapy) organized journals based on mission statements and do not represent an absolute classification of journal content. Alternative configurations of journal titles may be possible based on the unifying characteristics of behavior analysis. In addition, individual articles, regardless of journal categorization, may have aligned with alternate categories. The categorization system nonetheless permits the contextualization of findings given the breadth of behavior analysis. Moreover, results suggest that the categorization system corresponds with observable patterns in data collection and assessment.

Given the extensive of use of psychometric instrumentation in research and practice, studies featuring dependent measures evaluated through statistical techniques (e.g., test-retest) met criteria for inclusion. Reviewers applied the reliability assessment codes when the study authors reported reliability coefficients of psychometric instruments within the research sample. This poses potential problems due to (a) the potential incongruence between psychometric and behavioral properties of reliability and (b) conventions that hold instrument reliability as an inherent dimension of instruments rather than an attribute of case-by-case measurement. Notwithstanding the broadly defined category of reliability applied in the current review, however, few studies received the reliability code. The rejection of second hand reliability coefficients remains appropriate given routine use of psychometric instruments beyond populations featured in the initial test samples.

Future Directions for Researchers

Behavioral researchers of today have inherited a distinction for rigorous attention to the measurement and analysis of observable phenomenon (Johnston & Pennypacker, 2009; Matthews, 1998; Skinner, 1953). Continuing to deserve the reputation of our forbearers will require dissatisfaction with convention for the sake of convention. Dependent measure assessment should appear prominently in published behavioral research moving forward. The recent increasing trend of reporting measurement assessment data represents an obvious step. However, the use of IOA and its inherent limitations (Mudford et al., 2011) poses an additional challenge for researchers.

Moving toward reliability and accuracy measures presents data more firmly linked to natural phenomena contributing to a more robust behavioral research base. Accuracy of measures plays such a pivotal role that established standards guide fields such as Physics and Engineering (Dieck, 2014). Currently, many researchers already collect data in ways that facilitate both reliability and accuracy (i.e., permanent product and automated recording). Simply conducting reliability and/or accuracy procedures using extant data sources would immediately improve the quality of the data within the behavior sciences. Given the prominence of direct observation and IOA within the field an immediate change does not appear tenable. The continued reliance on direct observation, as opposed to more precise forms of measurement, likely stems from understandable limitations applied researchers face in terms of technology, time, and funds. In addition, measurement assessment represents one aspect contributing to the quality of evidence; studies with questionable internal validity, for example, provide little value to the field regardless of measurement quality. Nonetheless, researchers should continue to build upon areas of research predicated on human observation by (a) augmenting observational methods with permanent products (e.g., A/V representation) and (b) using novel approaches to automated data collection (e.g., Goodwin,

Intille, Albinali, & Velicer, 2011). In both cases, researchers can access both intraobserver agreement (i.e., reliability) and accuracy coefficients to compliment IOA measures.

References

Andersen, E., Jørgensen, L. G., & Hededam, L. T. (1990). Evans' classification of trochanteric fractures: An assessment of the interobserver and intraobserver reliability. *Injury, 21,* 377–378. http://dx.doi.org/10.1016/0020-1383(90)90123-C

Ayres, K., & Gast, D. L. (2010). Dependent measures and measurement procedures. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (129–165). New York, NY: Routledge.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1,* 91–97. http://dx.doi.org/10.1901/jaba.1968.1-91

Boyce, T. E., Carter, N., & Neboschick, H. (2000). An evaluation of intraobserver reliability versus interobserver agreement. *European Journal of Behavior Analysis, 1,* 107–114. http://dx.doi.org/10.1080/15021149.2000.11434159

Boykin, R. A., & Nelson, R. O. (1981). The effects of instructions and calculation procedures on observers' accuracy, agreement, and calculation correctness. *Journal of Applied Behavior Analysis, 14,* 479–489. http://dx.doi.org/10.1901/jaba.1981.14-479

Bryington, A. A., Palmer, D. J., & Watkins, M. W. (2002). The estimation of interobserver agreement in behavioral assessment. *Behavior Analyst Today, 3,* 323–328. http://dx.doi.org/10.1037/h0099978

Buekens, F., & Truyen, F. (2014). The truth about accuracy. In C. Martini & M. Boumans (Eds.), *Experts and consensus in social science* (pp. 213–229). Cham, Switzerland: Springer International Publishing.

Carr, J. E., & Britton, L. N. (2003). Citation trends of applied journals in behavioral psychology 1981–2000. *Journal of Applied Behavior Analysis, 36,* 113–117. http://dx.doi.org/10.1901/jaba.2003.36-113

Clarke, W. L., Cox, D., Gonder-Frederick, L. A., Carter, W., & Pohl, S. L. (1987). Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care, 10,* 622–628. http://dx.doi.org/10.2337/diacare.10.5.622

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8,* 411–426. http://dx.doi.org/10.1016/S0005-7894(77)80077-4

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Columbus, OH: Pearson.

Critchfield, T. S. (2002). Evaluating the function of applied behavior analysis a bibliometric analysis. *Journal of Applied Behavior Analysis, 35,* 423–426. http://dx.doi.org/10.1901/jaba.2002.35-423

Crowley-Koch, B. J., & Van Houten, R. (2013). Automated measurement in applied behavior analysis: A review. *Behavioral Interventions, 28,* 225–240. http://dx.doi.org/10.1002/bin.1366

Deitz, S. M. (1988). Another's view of observer agreement and observer accuracy. *Journal of Applied Behavior Analysis, 21,* 113. http://dx.doi.org/10.1901/jaba.1988.21-113

Dieck, R. H. (2014). Measurement accuracy. In J. G. Webster & H. Eren (Eds.), *Measurement, instrumentation, and sensors handbook* (pp. 5.1–5.14). New York, NY: Taylor & Francis. http://dx.doi.org/10.1201/b15474-7

Dixon, M. R., Nastally, B. L., Jackson, J. E., & Habib, R. (2009). Altering the near-miss effect in slot machine gamblers. *Journal of Applied Behavior Analysis, 42,* 913–918. http://dx.doi.org/10.1901/jaba.2009.42-913

Dorsey, L. B., Nelson, R. O., & Hayes, S. C. (1986). The effects of code complexity and of behavioral frequency on observer accuracy and interobserver agreement. *Behavioral Assessment, 8,* 349–363.

Farrance, I., & Frenkel, R. (2012). Uncertainty of measurement: A review of the rules for calculating uncertainty components through functional relationships. *The Clinical Biochemist Reviews/Australian Association of Clinical Biochemists, 33,* 49–75.

Fradenburg, L. A., Harrison, R. J., & Baer, D. M. (1995). The effect of some environmental factors on interobserver agreement. *Research in Developmental Disabilities, 16,* 425–437. http://dx.doi.org/10.1016/0891-4222(95)00028-3

Goodwin, M. S., Intille, S. S., Albinali, F., & Velicer, W. F. (2011). Automated detection of stereotypical motor movements. *Journal of Autism and Developmental Disorders, 41,* 770–782. http://dx.doi.org/10.1007/s10803-010-1102-z

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165–179. http://dx.doi.org/10.1177/001440290507100203

Hudson, S. S., Lewis, T., Stichter, J. P., & Johnson, N. W. (2011). Putting quality indicators to the test: An examination of 30 years of research. *Journal of Emotional and Behavioral Disorders, 19,* 143–155. http://dx.doi.org/10.1177/1063426610364986

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Joint Committee for Guides in Metrology. (2008). *Evaluation of measurement data—guide to the ex-*

pression of uncertainty in measurement. Sevres, France: Author.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10,* 141–150. http://dx.doi.org/10.1901/jaba.1977.10-141

Kazdin, A. E. (2012). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kelly, M. B. (1977). A review of the observational data-collection and reliability procedures reported in *The Journal of Applied Behavior Analysis*. *Journal of Applied Behavior Analysis, 10,* 97–101. http://dx.doi.org/10.1901/jaba.1977.10-97

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson.

Kubina, R. M., Jr., Kostewicz, D. E., & Datchuk, S. M. (2008). An initial survey of fractional graph and table area in behavioral journals. *The Behavior Analyst, 31,* 61–66.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher, 43,* 304–316. http://dx.doi.org/10.3102/0013189X14545513

Matthews, W. J. (1998). Let's get real: The fallacy of post-modernism. *Journal of Theoretical and Philosophical Psychology, 18,* 16–32. http://dx.doi.org/10.1037/h0091169

McDermott, P. A. (1988). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology, 3,* 225–240. http://dx.doi.org/10.1037/h0090563

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376–390. http://dx.doi.org/10.1037/0033-2909.86.2.376

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis, 42,* 165–169. http://dx.doi.org/10.1901/jaba.2009.42-165

Mudford, O. C., Zeleny, J. R., Fisher, W. W., Klum, M. E., & Owen, T. M. (2011). Calibration of observational measurement of rate of responding. *Journal of Applied Behavior Analysis, 44,* 571–586. http://dx.doi.org/10.1901/jaba.2011.44-571

Neely, L., Davis, H., Davis, J., & Rispoli, M. (2015). Review of reliability and treatment integrity trends in autism-focused research. *Research in Autism Spectrum Disorders, 9,* 1–12. http://dx.doi.org/10.1016/j.rasd.2014.09.011

Rapp, J. T., Carroll, R. A., Stangeland, L., Swanson, G., & Higgins, W. J. (2011). A comparison of reliability measures for continuous and discontinuous recording methods: Inflated agreement scores with partial interval recording and momentary time sampling for duration events. *Behavior Modification, 35,* 389–402. http://dx.doi.org/10.1177/0145445511405512

Skinner, B. F. (1939). *The behavior of organisms*. Acton, MA: Copley Publishing Group.

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Simon & Schuster.

Sturmey, P. (1994). Assessing the functions of aberrant behaviors: A review of psychometric instruments. *Journal of Autism and Developmental Disorders, 24,* 293–304. http://dx.doi.org/10.1007/BF02172228

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10,* 343–366.

Tapp, J. (1995). *Multiple options for observation in experimental studies*. Nashville, TN: Vanderbilt Kennedy Center.

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10,* 205–212. http://dx.doi.org/10.1023/A:1012295615144

Weinrott, M. R., & Jones, R. R. (1984). Overt versus covert assessment of observer reliability. *Child Development, 55,* 1125–1137. http://dx.doi.org/10.2307/1130165

White, G. H., Farrance, I., & the AACB Uncertainty of Measurement Working Group. (2004). Uncertainty of measurement in quantitative medical testing: A laboratory implementation guide. *The Clinical Biochemist Reviews, 25,* S1–S24.

Wolery, M. (2004). Monitoring children's progress and intervention implementation. In M. McLean, M. Wolery, & D. B. Bailey, Jr., (Eds.), *Assessing infants and preschoolers with special needs* (3rd ed., pp. 545–584). Upper Saddle River, NJ: Pearson.

Wu, S. M., Whiteside, U., & Neighbors, C. (2007). Differences in inter-rater reliability and accuracy for a treatment adherence scale. *Cognitive Behaviour Therapy, 36,* 230–239. http://dx.doi.org/10.1080/16506070701584367

Received August 31, 2015
Revision received January 22, 2016
Accepted January 29, 2016 ■