Embracing Our Visual Inspection and Analysis Tradition: Graphing Interobserver Agreement Data

Kathleen Artman, Mark Wolery and Paul Yoder

Remedial and Special Education 2012 33: 71 originally published online 1 September 2010
DOI: 10.1177/0741932510381653

Hammill Institute on Disabilities

Remedial and Special Education
33(2) 71–77

DOI: 10.1177/0741932510381653
http://rase.sagepub.com
SAGE

Embracing Our Visual Inspection and Analysis Tradition: Graphing Interobserver Agreement Data

Kathleen Artman[1], Mark Wolery[2], and Paul Yoder[2]

Abstract

Most investigators using single-case experimental designs use interobserver agreement (IOA) checks to enhance the credibility of the collected data, and they report the results of those assessments using percentage of agreement estimates. An alternative is to graph both observers' records of the measured behavior on the primary study graphs. Such graphing leads to greater transparency and is advocated for five reasons: (a) to make explicit how IOA assessments were distributed across the study, (b) to ensure agreement estimates are reported at the level of the measured behavior of interest rather than a broader observational code, (c) to detect observer drift, (d) to detect the effect of observer expectations, and (e) to put the IOA data in a more suitable context for assessing the internal validity of the study by eliminating the need for an arbitrary agreement criterion.

Keywords

research methodology, single participant, research methodology, quantitative, change, innovation

Since its infancy, the field of applied behavior analysis has wrestled with how best to quantify and represent observations of human behavior. Few would argue against the need for precise measurement of both participants' and observers' behaviors. In fact, direct measurement of the agreement between two observers has always been critical to the conduct of behavioral research (Baer, Wolf, & Risley, 1968). When it comes to how best to identify, quantify, and report this agreement, however, experts historically have disagreed (Baer, 1977; Hartmann, 1977; Kratochwill & Wetzel, 1977). The debates surrounding these disagreements offered a range of suggestions for improving methods for assessing and reporting interobserver agreement (IOA) data. Some suggestions, such as calculating data separately for agreement on occurrence and nonoccurrence in interval recording systems, were accepted and continue to appear in research reports. However, nonoccurrence agreement has no application in event sampling. Other laudable suggestions have not become part of common practice. The purpose of this article is to revisit one such suggestion—the graphic presentation of both observers' data (primary observer and IOA observer) on a common plot—more than 25 years after its last appearance in *Journal of Applied Behavior Analysis* (*JABA*).

Several authors have described graphing the second observer's data on the same graph as the primary observer's data as a means of establishing IOA agreement and supporting the believability of the experimental effects (Birkimer & Brown, 1979; Cooper, Heron, & Heward, 1987, 2007; Hawkins & Dotson, 1975). Hawkins and Dotson (1975) illustrated the weaknesses of standard methods for calculating percentage of agreement in interval recording systems. They recommended graphing both observers' data on a common plot to control for these weaknesses and enhance the believability of a functional relation. Four years later, Birkimer and Brown (1979) advocated graphic representation of IOA agreement data with several additions. They advocated plotting both observers' data alongside a "disagreement range" centered at the average of the two scores. The believability of a functional relation would be established when there was no overlap across the disagreement ranges of adjacent experimental conditions.

Presenting IOA data graphically was criticized for adding complexity to the visual analysis of primary data and expense to the publication of reports (Kratochwill & Wetzel, 1977). Birkimer and Brown's (1979) suggestion to avoid overlap

[1]Ohio State University, Columbus, OH, USA
[2]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Kathleen Artman, Ohio State University, 807 Kinnear Road, Room 220, Columbus, OH 43212
Email: artman.5@osu.edu

between adjacent disagreement ranges was criticized for introducing an increased risk of Type II error (Hartmann & Gardner, 1979; Kratochwill, 1979). It was suggested that graphing IOA data was more useful as an ongoing judgment aid for researchers than as a tool for consumers (Kratochwill, 1979; Kratochwill & Wetzel, 1977). If ongoing graphic displays showed problems with agreement, researchers were to rework the study and refrain from publishing the data. According to detractors of graphing IOA data, a simple summary of agreement and the assurance of sound methodology were sufficient evidence for consumers of research.

Perhaps because of these criticisms, graphic presentation of IOA agreement data failed to flourish in the field's publications. A 1977 review of the literature found 1% of the 222 articles published to that time in *JABA* had plotted both observers' data on the same graph (Kelly, 1977). We conducted a hand search of *JABA* publications between 1977 and May 2009 to determine if the debate of the late 1970s had an effect on IOA reporting practices. A brief spike in graphic presentation of IOA data occurred between 1977 and 1984. Seven articles offered some indication of IOA agreement on the primary data displays. Of these, one indicated only the temporal arrangement of agreement checks by marking their occurrence on the abscissa (Gladstone & Spencer, 1977), two indicated the temporal arrangement of agreement checks and the total percentage agreement between the two observers (Connis, 1979; Tucker & Berry, 1980), and four presented both observers' data on the same graph (Luce, Delquardi, & Hall, 1980; Tertinger, Greene, & Lutzker, 1984; Van Houten & Nau, 1980; Van Houten & Rolider, 1984). Despite this initial use, no articles published in *JABA* in the past 25 years presented graphic depiction of IOA data. Based on our experience but absent actual data collection, graphing both observers' data does not occur in other journals publishing single-case experimental research.

Since these articles appeared, the field clearly has relied on summary statements and statistics of IOA such as mean and range of percentage agreement estimates. These methods seem to offer results—achievement of a quantitative criterion value—that consumers can evaluate (Kratochwill & Wetzel, 1977). However, the criteria for suitable agreement are arbitrary and based on collective history rather than experimental research (Kennedy, 2005). A large body of literature supports the claim that indices of agreement are influenced by a variety of factors including base rate, observer expectations, and complexity of codes (Barlow & Hersen, 1984; Hawkins & Dotson, 1975; Hopkins & Hermann, 1977; Kazdin, 1977; Repp, Dietz, Boles, Dietz, & Repp, 1976).

The purpose of this article is to present a case for assessing IOA data by graphing the second observer's data alongside the primary observer's data in single-participant studies. In this article, we present assumptions about assessing agreement appropriately, provide a rationale and describe the advantages associated with graphing IOA data, and address the disadvantages with recommendations.

Assumptions About IOA Data

Given the nature of data collection in single-case experimental research, collecting IOA data is the most appropriate means for demonstrating the extent to which variability among sessions and design phases is consistent across observers. Several factors increase the likelihood investigators can have confidence in the collected data. Well-developed response definitions and clearly specified data collection procedures are necessary. The more complex the data collection system, the greater the likelihood of errors; thus, establishing a balance between simplicity and complexity is necessary. Observers should be trained to high levels of agreement in contexts similar to those in the study prior to initiating formal data collection. Recalibrating observers throughout the study also is an acceptable practice, when possible.

Ideally, the observers will be naïve to the purpose of the study and will not be aware of which experimental condition is in effect at any given observation. Knowledge of the purpose and of the conditions increases the possibility of observers' expectations influencing their recording behavior. In many cases, however, this is not possible because of limited resources and the nature of the experimental conditions. Similarly, observers should not be aware of when IOA assessments are being conducted because such knowledge has been shown to change observers' behavior (Romanczyk, Kent, Diament, & O'Leary, 1973). Keeping observers unaware of when IOA is being assessed is not possible in many, if not most, studies. Credibility of the data also is enhanced by collecting frequent IOA data. Ideally, IOA assessments should occur for all behaviors of each participant during each study condition. Finally, investigators should use conservative methods for calculating their agreement estimates. Formulas for calculating IOA vary in stringency (Kennedy, 2005), so researchers should carefully match agreement formulas to their data sets. For example, summary-level agreement (in which the total number of responses for each observer is tallied, the smaller number is divided by the larger number, and the value is multiplied by 100) is less conservative than total point-by-point agreement for interval data.

Justification and Advantages for Graphing IOA Data

The recommendation to graph the dependent measure values of both observers (primary observer and IOA observer) is justified in a long-standing tradition and strength of applied behavior analysis: It promotes openness and transparency with readers because the data are not transformed into summary statistics. Parsonson and Baer (1992), when arguing

for graphing and visual analysis of the primary study data, wrote,

> In representing the actual data measured, graphs can and usually do transform those data as minimally as possible. In those paradigms of knowing wherein the measureable data under study are the reality to be understood (cf. Heshusius, 1982, for a presumably different paradigm), that is an obvious virtue. (p. 16)

Behavior analysts expect study data to be presented in a minimally transformed manner so they can make independent judgments about whether and to what extent data patterns shifted with experimental manipulations. Authors cannot expect favorable reviews of their studies if all they report is the number of data points collected in each condition, the mean of those data, and their ranges. Yet this is the usual reporting practice for IOA data; the data are transformed into summary statistics and statements, clouding the reader's ability to make independent judgments about important aspects of data. If, as recommended here, both observers' data are graphed, tremendous transparency and openness result. This transparency is seen in five important ways, each of which is an advantage over providing only summary statements and statistics about IOA data.

First, graphing both observers' data makes apparent how IOA assessments are distributed throughout the study. It shows (a) for whom, which participants, IOA data were collected; (b) the conditions (study phases) in which those assessments occurred; (c) when within those conditions the IOA sessions were conducted; and (d) how often across the course of the study IOA data were collected. In general, a commonly used guideline is to collect IOA data in at least 20% of the sessions for each participant in each condition, with 25% to 33% of sessions being preferred (Cooper et al., 2007).

To determine whether current IOA reporting practices make explicit how IOA assessments are distributed in studies, each study reported in the first three issues of Volume 41 (2008) of *JABA* and the first issue of Volume 42 (2009) were reviewed; the fourth issue of 2008 was not reviewed because it contained a number of studies using group as compared to single-participant experimental designs. A total of 60 studies were described, and 58 (97.8%) reported IOA estimates; the 2 studies not reporting agreement estimates used automated data collection devices (e.g., slot machines; Hoon, Dymond, Jackson, & Dixon, 2008; Johnson & Dixon, 2009). All studies reporting agreement data did so with percentage of agreement estimates, none used correlation coefficients, and one study reported both percentage of agreement and kappa (McIver, Brown, Pfeiffer, Dowda, & Pate, 2009). In terms of reporting the frequency with which IOA assessments were collected, 33 (55%) of the articles described the percentage of sessions for the study as a whole with no indication of whether or how the IOA assessments were distributed across conditions or participants. Some studies provided more precise reporting of the frequency and distribution of IOA assessments: 7 (11.7%) studies described the frequency of assessment by study condition, 7 (11.7%) described the frequency by participant, and 11 (18.3%) described it by participant and condition. In terms of reporting the results of IOA assessments, 29 (48.0%) studies reported it for the study as a whole as compared to by condition or participant, 3 (5.0%) reported the results by condition, 15 (25.0%) reported the results by participants, and 11 (18.3%) reported the results by participant and condition. These data suggest the current reporting procedures are less than transparent. A reader cannot determine for a majority of the articles when during a study the IOA assessments occurred, whether they occurred in each condition, and whether they occurred for each participant. Furthermore, readers cannot determine for which conditions or participants low agreement estimates exist. This uncertainty would be corrected by graphing both observers' records of the primary data.

Second, graphing both observers' data will ensure the secondary observer's scores are at the same level of measurement as the dependent measure or measures used to test the research questions. Behavior analysis, measurement practices, and technology have advanced, resulting in increased opportunities to use more complex measurement systems (Mudford, Taylor, & Martin, 2009; Thompson, Felce, & Symons, 2000). Occasionally investigators collect data on multiple behaviors or behavioral categories despite having a focused interest in one or two. If the IOA summary statements are presented only for the entire observational code, of which the behavior of interest is but one element, the extent to which observers agreed the primary (and graphed) behavior occurred is not known. In other cases, investigators may present data as the ratio between two distinct behaviors of interest (e.g., children's correct responses to teachers' questions). Agreement on the numerator and/or denominator of the ratio does not necessarily ensure agreement on the ratio (Yoder & Symons, 2010). If both observers' data are graphed using the same metric for the behavior of interest, then the extent to which observers' records are similar or different on that particular metric is made obvious. It is a safeguard against reporting IOA estimates at the code as compared to the specific behavior level.

Third, graphing both observers' data allows the investigator and readers to examine the IOA data for *observer drift*. Kazdin (1977) defined observer drift as "the tendency of observers to change the manner in which they apply the definitions of behavior over time" (p. 143). If one observer is drifting, then over time the difference in the values of both observers' records will become greater. This phenomenon is shown in Figure 1. For the data in Figure 1, IOA assessments occurred in 40.0% of the sessions and the agreement averaged 89.0% with a range of 80.8% to 98.0%. However, agreement for the six IOA assessments was (in order of collection) 98.0%,

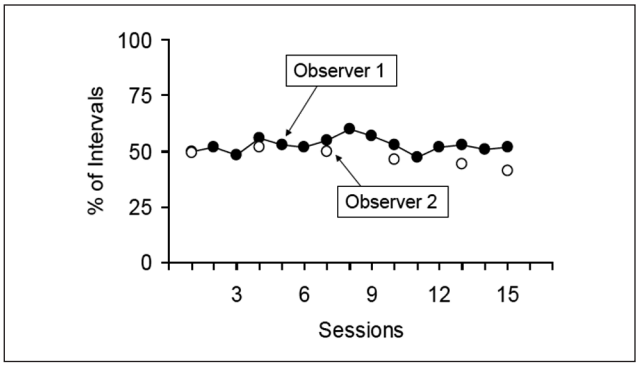

Figure 1. Hypothetical data illustrating how graphing both observers' data allows identification of observer drift

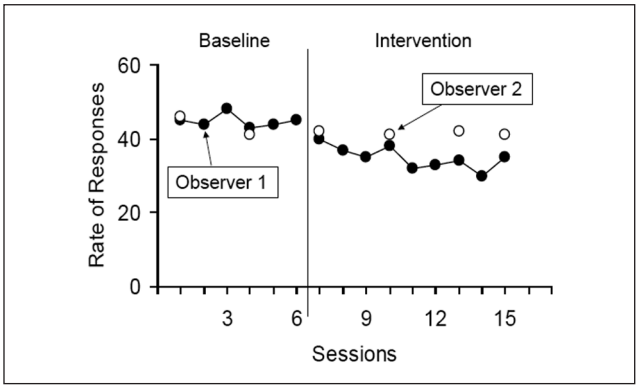

Figure 2. Hypothetical data illustrating how graphing both observers' data allows identification of observer expectations

94.6%, 90.9%, 86.8%, 83.0%, and 80.8%. Graphing does not indicate which observer is drifting, but it allows detection of the possibility that drift is occurring. When IOA data are presented in a summary statistic (mean percentage for a condition with a range [89.0%, range = 80.8% – 98.0%]), the ability to detect drift is lost, especially for the reader and potentially for all but attentive investigators.

Fourth, graphing both observers' data allows the investigator and readers to examine the IOA data for the effect of *observer expectations*. Kazdin (1977) described observer expectations by their effects: "Observers who look for behavior change are more likely to find it" (p. 147). If one observer has expectations for the behavior to change and the other does not, then this may be detected by graphing both observers' data. Such detection is most likely to be seen when experimental conditions change and observers are aware of the change, unavoidable events in some studies. This is illustrated in Figure 2, in which Observer 1 has expectations that the behavior will change (decrease) and Observer 2 is not influenced by such expectations. Observer 2's data remain quite stable across both conditions, whereas Observer 1's data show stable and flat levels during baseline and a decelerating trend during intervention. IOA data were collected in 33.3% of the baseline and 44.4% of the intervention sessions; the percentage of agreement in baseline was 96.6% (range = 95.4% to 97.8%) and in intervention was 88.6% (range = 81.0% to 95.2%). Calculating percentages of agreement does not allow the reader to detect the effects of observer expectations, but graphing might.

Fifth, graphing both observers' data allows readers to make better judgments about whether and to what extent lack of agreement (a type of instrumentation threat) compromises the internal validity of studies. The function of IOA data collection is to assess the extent to which both observers' records are sufficiently similar to warrant confidence in the primary study data. When the commonly reported percentage of agreement estimates are used, the question becomes, what level of agreement serves as a criterion for confidence? As Kazdin (1977) stated,

> Interobserver agreement is often considered adequate if it meets a prespecified level of agreement. Percentage agreement, one of the more commonly used measures, that reaches 70% or 80% often is considered satisfactory.[3] Yet, stressing the quantitative aspects of reliability ignores several assessment characteristics that dictate the meaningfulness of any agreement estimate. These include reactivity of reliability assessment, observer drift, complexity of the responses, information, expectancies, and feedback conveyed to the observers, and others (p. 142).

In the footnote, Kazdin went on to state,

> A quantitative criterion is difficult to invoke, in part, because the manner in which agreement should be defined and which of the available descriptive statistics for the computing reliability should be used are unresolved (Hartmann, 1976 [*sic*]; Hawkins and [*sic*] Dobes, 1975; Hawkins and [*sic*] Dotson, 1975.) The criterion for adequate agreement also depends on such diverse factors as variability and rate of the observed behaviors, the number of different response codes scored, and the strength of the intervention. (p. 142)

Kazdin's article was published more than 30 years ago, when the field had not resolved which descriptive statistics to use, although investigators currently consider 80% agreement as a minimum. Based on the data presented above, percentage of agreement has become the convention, with multiple calculations, depending on the type of recording system employed (Ayres & Gast, 2010; Kennedy, 2005). However, many of the issues Kazdin noted are no less problematic now than they were then, including identifying the "prespecified" criterion.

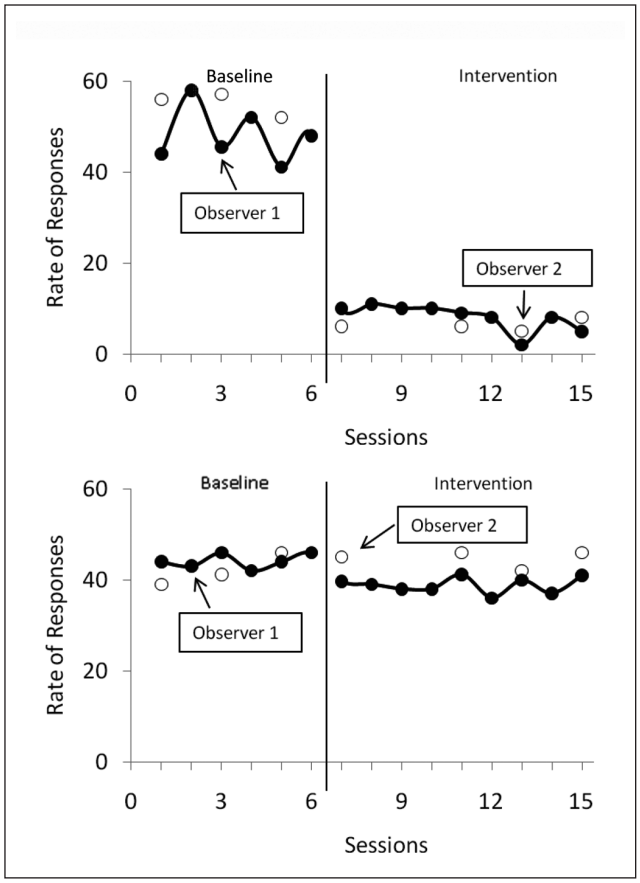

Figure 3. Hypothetical data illustrating how graphing both observers' data allows a judgment about behavior change despite low interobserver agreement percentages (top panel) and detection of no behavior change despite high interobserver agreement percentages (bottom panel)

Specifying an acceptable level of agreement remains elusive because it depends on multiple data characteristics and the conditions under which those data are gathered, as Kazdin indicated. Rather than ask what an acceptable level of agreement is before confidence in the data is warranted, the more functional question is, are the shifts in the data patterns across experimental conditions interpretable given the level of disagreement? This latter question is more easily answered when both observers' data are graphed. This is illustrated in the data displays shown in Figure 3. In the top panel, a clear and large shift in the level of data coincided with the introduction of the intervention condition, and the percentage of agreement estimates are low by usual conventions for baseline and intervention conditions. IOA was assessed in 50.0% of the baseline and 44.4% of the intervention sessions. Agreement for baseline was 79.1% (range = 78.6% to 79.7%); agreement for intervention was 56.9% (range = 40.0% to 66.7%). Thus, although the percentage of agreement would argue for not having confidence in the data, the graphed data suggest the rate of responding dropped substantially when the intervention condition was introduced. In the bottom panel of Figure 3, stable baseline data are followed by a small but consistent drop in level during the intervention condition. IOA data were collected in 50.0% of the baseline and 44.4% of the intervention sessions. During the baseline condition, agreement was 91.2% (range = 88.6% to 95.6%); during intervention, agreement was 90.5% (range = 88.1% to 95.0%). The percentages of agreement are well within commonly accepted criteria; however, graphing the IOA data calls into question whether a change occurred in the data during the intervention condition. The graphs in Figure 3 indicate the advantage of graphing the data over relying on arbitrary criterion levels for percentage of agreement. In addition, the example figures indicate that the size of the change and the percentage of nonoverlapping data affect the level of agreement required for readers to infer confidently a functional relation exists.

Graphing both observers' data clearly holds advantages over reporting summary statistics of percentage of agreement, but neither provides all the information needed to understand the factors influencing observers' behavior. These other factors include observers' knowledge of whether IOA is being assessed, the complexity or number of measured behaviors, and feedback given to observers (Kazdin, 1977). This information must be gleaned from the narrative of articles, and investigators are obliged to report them. Graphing both observers' data gives information about potential observer drift and expectancies, and it puts the extent of agreement or disagreement in the context of the data's variability and rate and of the intervention's strength.

Disadvantages of Graphing IOA Data and Recommendations

At least two potential disadvantages exist for graphing both observers' IOA data. First, adding the IOA observers' data to the graphs may increase the clutter in the graphs, which in turn may interfere with the visual analysis of the data (Kratochwill & Wetzel, 1977). Clearly, this is a possibility, but this issue can be settled as it should be, by conducting research to determine the influence of graphing IOA data on visual analysts' judgments. Thus, a clear recommendation is to conduct studies using different designs to evaluate the extent to which graphing IOA data alters visual analysts' judgments about shifts in data patterns. Such research should probably be done by design because graphing IOA observers' data may not detract from analysts' judgments for some designs but may for other designs (e.g., multielement designs).

The second disadvantage of graphing IOA data is it requires a change in research reporting behaviors in the face of a strong tradition of presenting the percentage of agreement as summary statements and statistics. Nonetheless, the advantages of graphing the IOA data appear to outweigh maintaining a

reporting practice simply because it has a long-standing tradition. Two recommendations ensue: First, in those cases in which all IOA sessions result in 100% agreement, graphing the data serves no useful function, and reporting them as a summary statement is warranted, assuming authors report precisely when and for whom the assessments occurred. Second, in cases in which the IOA sessions result in less than perfect agreement, then both observers' data should be graphed. If authors also wish to report the mean percentage of agreement and ranges, as is tradition, then no harm will necessarily occur. In fact, having both the graphed data and summary statements may illuminate the liabilities of relying on summary statements and statistics of percentage of agreement. The transparency resulting from graphing both observers' data argues strongly for adopting this practice.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

This publication was supported by a leadership personnel training grant from the Office of Special Education and Rehabilitation Services (OSERS) of the U.S. Department of Education. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of OSERS and the U.S. Department of Education.

References

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

Baer, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis*, *10*, 117–119.

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.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York, NY: Pergamon.

Birkimer, M. C., & Brown, J. H. (1979). A graphical judgmental aid which summarizes obtained and change reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis, 12*, 523–533.

Connis, R. T. (1979). The effects of sequential pictorial cues, self-recording, and praise on the job task sequencing of retarded adults. *Journal of Applied Behavior Analysis, 12,* 355–361.

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis.* Columbus, OH: Merrill.

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

Gladstone, B. W., & Spencer, C. J. (1977). The effects of modeling on the contingent praise of mental retardation counselors. *Journal of Applied Behavior Analysis, 10*, 75–84.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103–116.

Hartmann, D. P., & Gardner, W. (1979). On the not so recent invention of interobserver reliability statistics: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis, 12*, 559–560.

Hawkins, R. P., & Dobes, R. W. (1975). Behavioral definitions in applied behavior analysis: Explicit or implicit. In B. C. Etzel, J. M. LeBlanc, & D. M. Baer (Eds.), *New developments in behavioral research: Theory, methods, and applications. In honor of Sidney W. Bijou* (pp. 167–188). Hillsdale, NJ: Lawrence Erlbaum.

Hawkins, R. P., & Dotson, V. A. (1975). Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of inter-observer agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application* (pp. 359–376). Englewood Cliffs, NJ: Prentice Hall.

Heshusius, L. (1982). At the heart of the advocacy dilemma: A mechanistic world view. *Exceptional Children, 49*, 6–13.

Hoon, A., Dymond, S., Jackson, J. W., & Dixon, M. R. (2008). Contextual control of slot-machine gambling replication and extension. *Journal of Applied Behavior Analysis, 41*, 467–470.

Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10*, 121–126.

Johnson, T. E., & Dixon, M. R. (2009). Influencing children's pre-gambling game playing via conditional discrimination training. *Journal of Applied Behavior Analysis, 42*, 73–81.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10,* 141–150.

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.

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

Kratochwill, T. R. (1979). Just because it's reliable doesn't mean it's believable: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis, 12*, 553–557.

Kratochwill, T. R., & Wetzel, R. J. (1977). Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis, 10*, 133–139.

Luce, S. C., Delquardi, J., & Hall, R. V. (1980). Contingent exercise: A mild but powerful procedure for suppressing inappropriate verbal and aggressive behavior. *Journal of Applied Behavior Analysis, 13*, 583–594.

McIver, K. L., Brown, W. H., Pfeiffer, K. A., Dowda, M., & Pate, R. R. (2009). Assessing children's physical activity in their homes: The observational system for recording physical activity children-home. *Journal of Applied Behavior Analysis, 42*, 1–16.

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.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, the current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Lawrence Erlbaum.

Repp, R. C., Dietz, D. E. D., Boles, S. M., Dietz, S. M., & Repp, C. F. (1976). Technical article: Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis, 9*, 109–113.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. (1973). Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis, 6*, 175–184.

Tertinger, D. A., Greene, B. F., & Lutzker, J. R. (1984). Home safety: Development and validation of one component of an ecobehavioral treatment program for abused and neglected children. *Journal of Applied Behavior Analysis, 17*, 159–174.

Thompson, T., Felce, D., & Symons, F. (2000). *Behavioral observation: Technology and applications in developmental disabilities.* Baltimore, MD: Brookes.

Tucker, D. J., & Berry, G. W. (1980). Teaching severely multihandicapped students to put on their own hearing aids. *Journal of Applied Behavior Analysis, 13*, 65–75.

Van Houten, R., & Nau, P. A. (1980). A comparison of the effects of fixed and variable ratio schedules of reinforcement on the behavior of deaf children. *Journal of Applied Behavior Analysis, 13*, 13–21.

Van Houten, R., & Rolider, A. (1984). The use of response prevention to eliminate nocturnal thumbsucking. *Journal of Applied Behavior Analysis, 17*, 509–520.

Yoder, P. J., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer.

About the Authors

Kathleen Artman, PhD, is a postdoctoral research scientist in the College of Education and Human Ecology at Ohio State University. Her current interests include teacher training and professional development, positive behavior support, and research methodology.

Mark Wolery, PhD, is a professor of special education at Peabody College, Vanderbilt University. His research interests include transfer of stimulus control in inclusive preschool classrooms, children with autism, and single-participant research methods.

Paul Yoder, PhD, is a professor of special education at Vanderbilt University. His interests include observational measurement, research methodology, and social influences on communication and language development in young children with a variety of disorders.