Journal of
Applied Behavior Analysis



Journal of Applied Behavior Analysis 2022, **55,** 357–368 NUMBER 2 (SPRING)

# Interobserver agreement: A preliminary investigation into how much is enough?

Nicole L. Hausman

Department of Behavioral Psychology, Kennedy Krieger Institute and the Department of Psychiatry and Behavioral Sciences, the Johns Hopkins University School of Medicine

Noor Javed and Molly K. Bednar

Department of Behavioral Psychology, Kennedy Krieger Institute

Madeleine Guell

Department of Psychology, Johns Hopkins University

Erin Schaller

Little Leaves Behavioral Services

Rose E. Nevill

Department of Behavioral Psychology, Kennedy Krieger Institute and the Department of Psychiatry and Behavioral Sciences, the Johns Hopkins University School of Medicine

SungWoo Kahng

Department of Applied Psychology, Rutgers University

Interobserver agreement (IOA) is important for research and practice, and supports the consistency of behavioral data (Kahng et al., 2011). Although general parameters for how much IOA is needed have been suggested (Bailey & Burch, 2018), it is unknown if the total number of sessions with IOA might impact the IOA coefficient. In this study, IOA was reanalyzed using functional analysis data at various cutoffs. Obtained IOA from these analyses was then compared to the original IOA. Overall, results suggested that, at least when using highly trained observers in a structured clinical setting, there were no significant differences in IOA across cutoffs. However, IOA was sensitive to overall rate of responding in the functional analysis. These data are encouraging, particularly for practitioners, because they provide preliminary support that the amount of sessions with IOA may not be as important as the consistency of the data.

*Key words:* interobserver agreement, reliability, data collection

Data collection that is reliable and valid is critical to applied behavior analysis (Hartmann, 1977; Kazdin, 1977; Kennedy, 2005; Kratochwill & Wetzel, 1977; Watkins & Pacheco, 2000). Typically, trained observers directly observe an individual during behavioral sessions and collect data on the occurrence of the dependent variable (i.e., target behavior) within that observation period. Two observers simultaneously and independently collect data for a subset of sessions and an appropriate measure of interobserver agreement (IOA) is calculated to provide an estimate of observer consistency (Kahng et al., 2011). An IOA score of 80% or greater is generally considered acceptable by the behavior-analytic community, which lends confidence that changes in the

The authors wish to thank Michael Kranak for his comments on earlier versions of the manuscript.

Address correspondence to: Nicole Hausman, Department of Behavioral Psychology, Kennedy Krieger Institute and the Department of Psychiatry and Behavioral Sciences, the Johns Hopkins University School of Medicine. Email: drhausmanbcbad@fullspectrumaba.com

doi: 10.1002/jaba.811

dependent variables are a result of the intervention and not due to differences in data collection across observers (Page & Iwata, 1986).

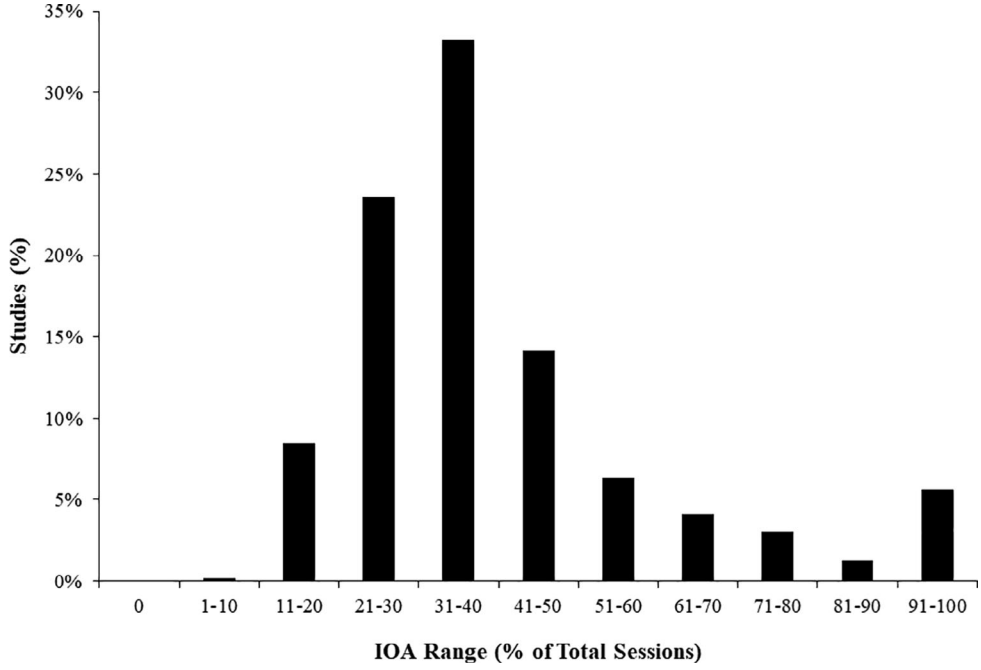
Although there are guidelines for selecting the most appropriate measure of IOA given certain parameters such as response rate or data collection method (e.g., Rolider et al., 2012), there is no empirical guidance on the amount of sessions needed to calculate IOA in order to demonstrate the data are reliable or consistent (i.e., what percentage of sessions should include a secondary observer for the purposes of calculating IOA). Variability in clinical guidelines pertaining to IOA may be particularly difficult to negotiate in clinical settings with limited resources. Assuming data collectors are trained on both the data collection system and the operational definitions of the target behavior, it is plausible that IOA should remain relatively stable regardless of how many sessions are used to calculate IOA. That is, we might expect that if certain basic requirements are met, IOA calculations may not be expected to be statistically different whether IOA is calculated based on 10% or 90% of sessions.

Guidelines on how many sessions should be used to calculate IOA varies between 20%–33% of sessions (Kennedy, 2005; Poling et al., 1995). Bailey and Burch (2018) suggested that IOA should be calculated for a minimum of 30% of sessions. Cooper et al. (2019) suggested that a higher percentage of data should have IOA for continuous measures, which accounted for a higher percentage of more than half of studies reporting data on free-operant behavior in the *Journal of Applied Behavior Analysis* (*JABA*) from 1995–2005 (Mudford et al., 2009). Taking these guidelines together, a minimum of 30% of sessions appears to be the most consistent recommendation. It should also be noted that these guidelines assume that individual IOA coefficients for the relevant dependent variables remain high, as even reporting 70% of sessions with low IOA coefficients calls the consistency of observers' data collection into question.

Despite these general guidelines for overall IOA, practical limitations (e.g., access to additional data collectors) may influence the actual percentage of sessions in which a second observer is available. These guidelines may not reflect actual clinical or research practices. For the purposes of the current study, the authors conducted a review of research studies published in *JABA* from the Spring 2015 to the Fall 2018 issue to evaluate recent conventions in reporting IOA. Although IOA is reported for the vast majority (98.88%) of studies in *JABA* from 2014-2018, there is a high degree of variability across studies in the proportion of sessions during which two observers independently collected data for the purposes of calculating IOA ($M$ = 41%, range: 3%–100% of sessions; Figure 1). The majority of studies reported IOA between 11%–50% of sessions, with 33% of studies reporting between 31%–40% of sessions with IOA. These data suggested that there seems to be adherence to general guidelines that IOA be calculated for 20%–33% of total sessions; however, the overall percentage of IOA reported in the literature varies.

Despite the existence of general practice guidelines suggesting IOA should be calculated based on 30% or more sessions, it does not appear as though this percentage was empirically derived (Kennedy, 2005). To date, there exists no empirical support for how many sessions should be used to calculate IOA, assuming individual IOA coefficients are acceptable. Therefore, the purpose of the current study was to provide preliminary empirical support to help guide clinicians and researchers in determining the most appropriate number of sessions needed to calculate IOA. We did this by recalculating percentages of sessions used to calculate IOA from functional analysis data collected by highly trained observers in a structured clinical environment. A secondary purpose was to determine if frequency of problem behavior predicted improvements or reductions in IOA across various calculation methods.

**Figure 1**
*Distribution of Ranges of Overall IOA from JABA Publications (2014-2018)*




## Method

### Participants and Setting

Functional analysis data (FA; Iwata et al., 1994) from inpatients admitted to a specialized unit for the assessment and treatment of severe problem behavior between 2013 and 2018 were selected for inclusion in this study. Specifically, we obtained multielement FA data for each inpatient ($n$ = 100), as this was a common behavioral assessment procedure that was conducted with all inpatients. Most participants (86%) were between 5–17 years of age at admission and 84% were male. Additionally, 90% were diagnosed with autism spectrum disorder and 82% with intellectual disability.

### Measures

Five parameters of the number of sessions with IOA were calculated for each FA. *Original IOA* was defined as the percentage of sessions with IOA initially reported by the clinical team for the FA. Subsequently, we recalculated various percentages of IOA between 10%-30% of total sessions. Therefore, *30% IOA* was defined as 30% of total sessions with IOA, *25% IOA* was defined as 25% of total sessions with IOA, *15% IOA* was defined as 15% of total sessions with IOA, and *10% IOA* was defined as 10% of total sessions with IOA (rounding procedures for specific cases are defined in further detail below). All measures were calculated by dividing the number of sessions with IOA by the total number of sessions conducted in the FA and multiplied by 100.

### Procedures

#### *FA Selection Criteria*

The following inclusion criteria were utilized for choosing the FA for each case: (a) a multielement design was used, (b) the target response occurred across at least three sessions at a rate of 0.3 responses per minute (RPM) or

greater, and (c) a primary and secondary observer collected data for at least 30% of total sessions (i.e., data were collected by two independent observers for a minimum of 30% of total sessions). The FA could be conducted at any time during the participant's inpatient admission, but was typically initiated within the first 2 weeks. Clinical records were reviewed from the last 10 years and cases included in these analyses if they met the above inclusion criteria until a target sample size of 100 was reached.

### *Selection of Target Behavior*

Upon admission, the patient's clinical team selected the primary behaviors (e.g., aggression, self-injurious behavior [SIB], disruption) to target in the FA and in subsequent treatment. The current study evaluated IOA for the same target behavior identified by the clinical team. If the clinical team selected more than one behavior to target in the FA or narrowed the definition of the primary behavior into different topographies (e.g., SIB was further separated into head-directed, body-directed, and self-biting), the behavior or topography that occurred at moderate-to-high rates (i.e., at or above 0.3 RPM) during the FA was selected (e.g., head-directed SIB).

### *Training in Data Collection*

Each patient's clinical team was composed of three individuals, a primary therapist and two backup therapists, who rotated primary and reliability data collection responsibilities. Clinical team members were typically individuals seeking an advanced degree in applied behavior analysis or a related field, and/or had experience in behavioral assessment procedures and collecting data for individuals with severe problem behavior. Each clinical team member received prior training on the data collection system (B-DataPro; Bullock et al., 2017) and had to demonstrate an ability to collect reliable data for any individual inpatient before formally using their data for clinical purposes. Specifically, reliable data collectors for a patient were those clinical team members that demonstrated IOA of 90% or greater for three consecutive sessions when compared to an already reliable data collector. This process was repeated for each new patient with whom the data collector worked. That is, data collector reliability was retested for multiple patients. IOA calculations did not include training data (i.e., data obtained prior to meeting criterion). It is important to note that all data collectors had previous experience in the assessment and treatment of severe problem behavior, had specific training in data collection systems, used standardized key assignments in B-DataPro (when applicable; e.g., key 1 for SIB, key 8 for prompt, "R" for reinforcement interval onset/offset), and standardized session descriptions for the FA. Thus, all data collectors had experience in data collection and all data were collected in a highly structured clinical setting.

### *Manipulation of IOA*

The total number of sessions with IOA for each participant was subsequently.

manipulated such that 30%, 25%, 15%, and 10% IOA (i.e., programmed IOA percentages) could be calculated for each participant. Recall that data collected by two independent observers for a minimum of 30% of sessions was one of the inclusion criteria for this study. In order to manipulate the percentage of sessions with IOA, sessions with IOA were deleted using a random number generator until 30%, 25%, 15%, and 10% IOA were obtained. We randomly determined which sessions with IOA would be excluded and then recalculated IOA for the FA without those sessions. For example, if 40 sessions were conducted for an FA and the original percentage of sessions with IOA was 50% (i.e., there were 20 sessions with IOA), then IOA data for 14, 15, 17, and 18 sessions would be randomly selected and deleted to obtain 30%, 25%, 15%, and 10% IOA,

respectively. If the random number generator provided a session number for which IOA was not collected, then the next random session number generated, which had IOA, was deleted.

In many cases, it was mathematically impossible to obtain the exact programed IOA percentages targeted in the simulation. Therefore, a $\pm$ 0.50% window was used. For example, if there were 31 sessions conducted in a FA, it was impossible to calculate 10% IOA. However, it was possible to calculate 9.68% of sessions with IOA, which was used for the 10% IOA simulation. For six participants, IOA of less than 9.50% was calculated for the 10% IOA simulations. Additionally, 25% IOA could not be calculated for Participant 54 because it was mathematically impossible to calculate a percentage near 25% without those percentages falling in the range of 15% or 30% IOA due to the total number of sessions conducted in the FA.

Exact (Repp et al., 1976), partial (Mudford et al., 2009), total (Bijou et al., 1968), occurrence, and nonoccurrence (Harris & Lahey, 1978) agreement-within-intervals methods were then calculated for the original IOA and programed percentages (i.e., 30%, 25%, 15%, 10%; see Table 1 for calculations). Note that total agreement for the purposes of this study was calculated differently than Total Count agreement as described by Cooper et al. (2019; Table 1). Each session was divided into consecutive 10-s bins and the respective agreement coefficient was calculated using B-DataPro software (Bullock et al., 2017). Note that for partial-interval agreement, a score of 1.0 was assigned for intervals during which both reviewers recorded that no behavior occurred. These IOA calculations were reported because they represent various calculation methods that may be reported by clinicians or researchers, and may be differentially impacted by response rate and number of observations. Thus, calculating these various coefficients allowed for a more complete evaluation of how commonly used IOA coefficients may vary by manipulating the total number of sessions in the analysis.

The Statistical Package of the Social Sciences (IBM Corporation, 2018) computer software compared differences in IOA across different programmed percentages. Analysis of variance (ANOVA) was conducted to determine if programmed percentages of IOA (original, 30%, 25%, 15%, 10%) resulted in significant differences across IOA types. Next, we examined the extent to which response rate was associated with higher IOA across cutoff levels and IOA types. Finally, simple linear regressions were calculated to examine whether lower response rate of problem behavior predicted higher IOA across different IOA coefficients. Power analysis was based on our most complex analysis (a repeated measures ANOVA with one group and five measurements) and conducted in G*Power to identify a sufficient sample size. The desired sample size was 88 (Faul et al., 2013) given an alpha of .05 and a power of 0.80 to detect a medium effect size ($f$ = 0.25).

## Results

### Original Data

#### *Interobserver Agreement*

In these clinical data sets, IOA was reported for an average of 60.84% (range: 32%–100%) of sessions across all multielement FAs. Recall that these IOA calculations were independently calculated and reported by each patient's clinical team during the course of their admission. Agreement was generally high across all calculation methods at an average of 95.82% (range: 77.22%–100%) for exact, 97.39% (range: 87.73%–100%) for partial-interval, 98.39% (range: 92.13%–100%) for total, 76.07% (range: 0%–100%) for occurrence, and 97.51% (range: 86.42%–100%) for nonoccurrence agreement.

#### *Rate Data*

As IOA calculations can be sensitive to response rate, the average rate of responding during each FA condition was calculated

**Table 1**

*IOA Calculations*

| Type | Included Intervals | Agreement | Formula |
|---|---|---|---|
| Exact Agreement (EA) | All | Both observers same # responses | $\left(\frac{Agreements}{Agreements + Disagreements}\right) x\,100$ |
| Partial Interval Agreement (PIA) | All | Both observers score ≥ 1 or 0 responses | $\left(\frac{\sum\left(\frac{Smaller}{Larger}\right)}{Total\ Intervals}\right) x\,100$ |
| Total Agreement (TA) | All | Both observers score ≥ 1 or 0 responses | $\left(\frac{Agreements}{Agreements + Disagreements}\right) x\,100$ |
| Occurrence Agreement (OA) | Only intervals with at least one observer scoring ≥1 recorded responses | Both observers score ≥ 1 responses | $\left(\frac{Agreements}{Agreements + Disagreements}\right) x\,100$ |
| Nonoccurrence Agreement (NOA) | Only intervals with at least one observer scoring 0 responses | Both observers score 0 responses | $\left(\frac{Agreements}{Agreements + Disagreements}\right) x\,100$ |

*Note*. IOA calculation methods within 10 s intervals. Please see Bullock et al. (2017) for a full review of these calculations and the B-DataPro program.

(Figure 2). Rates for each individual are also included to highlight the variability in response rate in the sample across conditions. During the alone and ignore conditions, the average rate of responding was 2.68 RPM (range: 0–41.16). During the attention and divided attention conditions, average rate of responding was 1.95 RPM (range: 0–14.40). Average rate of responding during the contingent tangible conditions averaged 1.19 RPM (range: 0–8.44). During the escape conditions (escape from demands, noise, and social interaction), average rate of responding was 0.94 RPM (range: 0–15.60). For one individual, an adult compliance with child mands test condition was added ($M$ = 0.99 RPM). Finally, for the toy play condition, responding averaged 0.97 RPM (range: 0–33.95). Therefore, all conditions included sessions during which there were zero to low rates of responding and included some sessions during which there were high rates of responding.

## IOA Across Types of Calculations

For all IOA coefficients, each session was divided into consecutive 10-s bins and the respective agreement coefficient was then subsequently calculated using B-DataPro software (Bullock et al., 2017; Table 1).

### *Exact Agreement*

For original data, exact agreement averaged 95.82% (range: 70.83%–100%). Average exact agreement varied minimally across IOA cutoffs, averaging 95.35% (range: 70.83%–100%), 95.60% (range: 58.67%–100%), 95.27% (range: 58.67%–100%), and 95.49% (range: 51.72%–100%) for the 30%, 25%, 15%, and 10% cutoffs, respectively. It should be noted that although the average recalculated exact agreement scores did not vary significantly, wider ranges of individual IOA scores were obtained for the 15% and 10% cutoffs (Figure 3).

### *Partial-Interval Agreement*

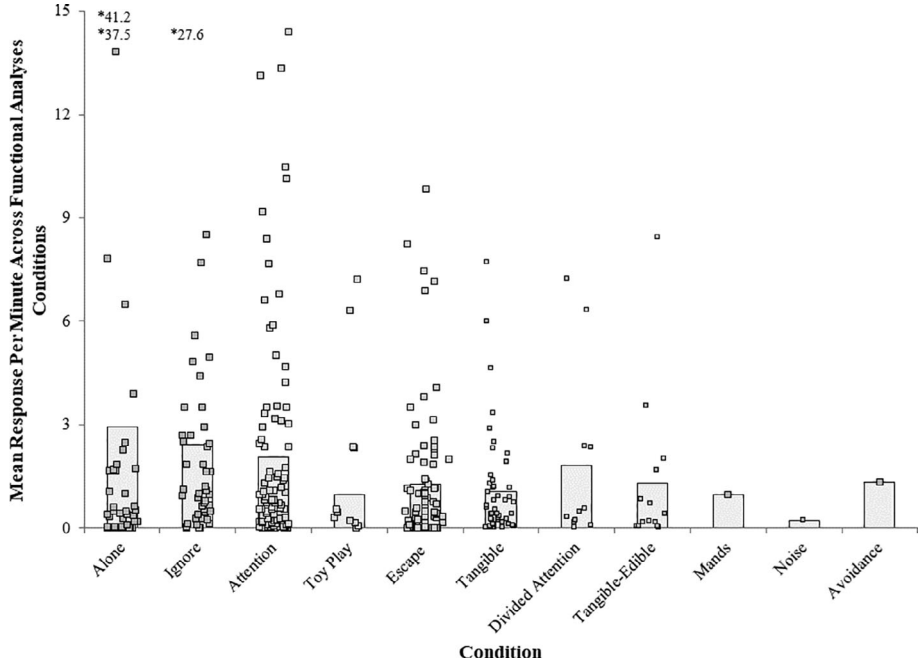
Partial-interval agreement averaged 97.39% (range: 87.73–97.76%) from the original sample data. Average agreement also varied minimally across the recalculated IOA cutoff points, with means of 97.23% (range: 80.33%–100%), 97.36% (range: 85.23%–100%), 96.98% (range: 71.90%–100%), and 97.31% (range: 76.39%–100%) across 30%, 25%, 15%, and 10% cutoffs, respectively (Figure 3).

### *Total Agreement*

Total agreement averaged 98.39% (range: 92.13%–98.54%) in the original sample. For the 30%, 25%, 15%, and 10% cutoffs,

**Figure 2**

*Individual and Mean Response Rates across Functional Analysis Conditions*




*Note*. Mean (bars) and distribution of individual (squares) responses per minute (RPM) during the functional analysis across conditions ($N$ = 100).

agreement averaged 98.33% (range: 90.42%–100%), 98.48% (range: 94.57%–100%), 98.13% (range: 86.67%–100%), and 98.40% (range: 80.56%–100%), respectively (Figure 3).

### *Occurrence Agreement*

Occurrence agreement from the original sample averaged 76.07% (range: 0%–100%; Figure 3). For the 30% cutoff, IOA averaged 78.25% (range: 23.02%–100%). For the 25% cutoff, IOA averaged 78.27% (range: 0%-100%). For the 15% cutoff, IOA averaged 72.36% (range: 0%–100%). Finally, for the 10% cutoff, IOA averaged 77.54% (range: 0%–100%). It should be noted that the variability in IOA scores was highest for OA in the original sample and across all recalculated IOA cutoffs, relative to other IOA calculation methods. Additionally, none of the mean occurrence agreement scores were acceptable, with all average scores falling below 80%.

### *Nonoccurrence Agreement (NOA)*

Nonoccurrence agreement for the original sample averaged 97.51% (range: 86.42%–100%). Across recalculated cutoff scores, IOA averaged 97.27% (range: 78.28%–100%), 97.56% (range: 89.38%–100%), 96.99% (range: 68.83%–100%), and 97.43% (range: 66.67%–100%) for the 30%, 25%, 15%, and 10% cutoffs, respectively (Figure 3).

## IOA Across Session Cutoffs

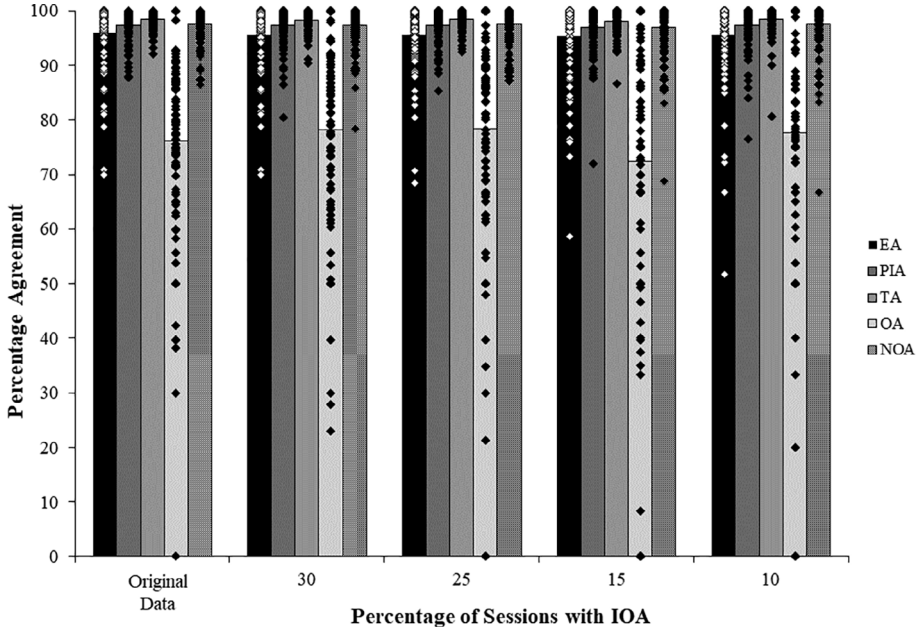
Differences in IOA across different sessions were compared via ANOVA. There were no significant differences in IOA when a minimum of 30%, 25%, 15%, and 10% of total sessions with reliability were used (Table 2).

## IOA and Response Rate

Examining the relationship between response rate and IOA types across programmed cutoffs

**Figure 3**

*Change in IOA across Reanalyzed Datasets*




*Note*. Average IOA for individual FAs are depicted by the data points for each cutoff and each IOA calculation type.

(original reliability, 30%, 25%, 15%, 10%), all IOA coefficient scores with the exception of occurrence agreement were significantly negatively correlated with response rate across all reliability cutoffs (Table 3). Strength of correlations was highest for original IOA level and decreased as IOA coefficients decreased from 30% to 10%. Correlations between response rate and IOA were large when using exact agreement across different percentage cutoffs, medium to large when using partial-interval agreement, small to medium when using total agreement, small to large when using nonoccurrence agreement, and small to medium when using partial interval agreement. Due to the significant correlations between response

**Table 2**

*Descriptive Statistics for IOA Attained Across Different Reliability Cutoffs*

| IOA | 30 | 25 | 15 | 10 | $F(3)$ | $p$ | $\eta_p^2$ |
|---|---|---|---|---|---|---|---|
| EA | 95.51 (5.83) | 95.58 (5.87) | 95.27 (6.99) | 95.47 (7.72) | 0.13 | 0.94 | 0.001 |
| PIA | 97.22 (3.31) | 97.35 (3.01) | 96.98 (4.12) | 97.30 (3.98) | 0.61 | 0.612 | 0.006 |
| TA | 98.32 (1.78) | 98.48 (1.62) | 98.13 (2.26) | 98.38 (2.61) | 0.95 | 0.419 | 0.01 |
| OA | 79.10 (15.44) | 79.27 (17.04) | 74.81 (21.92) | 76.81 (23.76) | 1.61 | 0.188 | 0.022 |
| NOA | 97.25 (3.64) | 97.55 (3.09) | 96.99 (4.70) | 97.40 (4.80) | 0.75 | 0.523 | 0.008 |

*Note*. Means and standard deviations (parentheses) for each IOA calculation across cutoffs.

**Table 3**

*Correlations Between Response Rate and IOA Type Across Programmed Cutoffs*

| IOA | Original | 30% | 25% | 15% | 10% |
|---|---|---|---|---|---|
| EA | **-.663**$^{**}$ | **-.645**$^{**}$ | **-.729**$^{**}$ | **-.621**$^{**}$ | **-.564**$^{**}$ |
| PIA | **-.557**$^{**}$ | **-.480**$^{**}$ | **-.538**$^{**}$ | **-.461**$^{**}$ | **-.433**$^{**}$ |
| TA | **-.353**$^{**}$ | **-.300**$^{**}$ | **-.302**$^{**}$ | **-.280**$^{**}$ | **-.211**$^{*}$ |
| OA | 0.195 | 0.170 | 0.158 | 0.150 | 0.178 |
| NOA | **-.486**$^{**}$ | **-.441**$^{**}$ | **-.511**$^{**}$ | **-.424**$^{**}$ | **-.289**$^{**}$ |

$^{*}p < .05$.
$^{**}p < .01$.
$^{***}p < .001$.

rate and IOA, it was of interest to determine the amount of variability in IOA that was predicted by response rate in problem behavior across different IOA coefficients, anticipating that lower response rate would predict higher IOA across coefficients. Response rate significantly predicted variability across all IOA coefficients except for occurrence agreement. Across sessions with IOA cutoffs, response rate predicted the highest amount of variability for exact, followed by partial-interval, and nonoccurrence agreement. Response rate predicted lowest amount of variability in total agreement. At the highest cutoff of 30% sessions with IOA, 42% of variability in exact agreement was predicted by response rate. At 10% sessions with IOA, 32% of variability in exact agreement was predicted by response rate (Table 4).

**Table 4**

*Simple Linear Regression for IOA Types Across Reliability Cutoffs Predicted by Response Rate*

| IOA | $R^2$ | *df* | *F* | *p* |
|---|---|---|---|---|
| Original EA | 0.44 | 2,98 | 37.58 | *** |
| Original PIA | 0.31 | 2,98 | 21.65 | *** |
| Original TA | 0.13 | 2,98 | 6.90 | ** |
| Original OA | 0.04 | 2,98 | 1.95 | 0.148 |
| Original NOA | 0.24 | 2,98 | 15.12 | *** |
| 30% EA | 0.42 | 1,99 | 69.77 | *** |
| 30% PIA | 0.23 | 1,99 | 29.30 | *** |
| 30% TA | 0.09 | 1,99 | 9.69 | ** |
| 30% OA | 0.03 | 1,99 | 2.83 | 0.096 |
| 30% NOA | 0.19 | 1,99 | 23.62 | *** |
| 25% EA | 0.53 | 1,99 | 111.28 | *** |
| 25% PIA | 0.29 | 1,99 | 40.00 | *** |
| 25% TA | 0.09 | 1,99 | 9.87 | ** |
| 25% OA | 0.03 | 1,99 | 2.45 | 0.121 |
| 25% NOA | 0.26 | 1,99 | 34.62 | *** |
| 15% EA | 0.39 | 1,98 | 60.99 | *** |
| 15% PIA | 0.21 | 1,98 | 26.17 | *** |
| 15% TA | 0.08 | 1,98 | 8.27 | ** |
| 15% OA | 0.02 | 1,98 | 2.06 | 0.155 |
| 15% NOA | 0.18 | 1,98 | 21.32 | *** |
| 10% EA | 0.32 | 1,97 | 45.60 | *** |
| 10% PIA | 0.19 | 1,97 | 22.61 | *** |
| 10% TA | 0.04 | 1,97 | 4.57 | * |
| 10% OA | 0.03 | 1,97 | 2.55 | 0.114 |
| 10% NOA | 0.08 | 1,97 | 8.96 | ** |

$^{*}p < .05$.
$^{**}p < .01$.
$^{***}p < .001$.

## Discussion

Results from the recalculated IOA coefficients suggested that no significant differences in IOA were obtained at the various total IOA cutoffs (i.e., 30%, 25%, 15%, and 10%) with FA data collected by highly trained observers. However, the simulated IOA scores were sensitive to response rate. As anticipated, response rate positively predicted greater variability in IOA, with higher response rate leading to increased variability. Response rate predicted the greatest amount of variability in exact agreement, regardless of IOA cutoff. These results suggest that although there were no differences in the overall recalculated IOA at the various cutoffs when compared to the original IOA scores, response rate for any individual case may influence obtained IOA, and these differences may vary depending on the type of IOA evaluated.

These results should be interpreted with caution as they were obtained by recalculating IOA using cutoffs based on a sample of FA data collected by highly trained observers in a structured, inpatient clinical setting. Although we found no statistically significant differences using a clinical sample of FA data obtained in a highly specialized facility, it is possible that recalculating IOA for other types of clinical data (e.g., correct responses) and collected in

other, potentially less controlled settings (e.g., schools, in homes) may differ more significantly. Therefore, these data represent only a starting point for determining how much IOA is "enough," and additional studies are warranted to establish empirically supported conventions for reporting IOA.

The interpretation of IOA is complex and based on a combination of the overall percentage of sessions where a second observer collected data, the appropriateness of the IOA calculation method used given the type of data collected, and the IOA coefficient itself. The relative impact of effect size and IOA is particularly relevant to this discussion, as studies that are reporting a greater effect size but lower overall IOA may be acceptable. On the other hand, studies with lower effect sizes may necessitate more IOA to support the consistency of the data collected (see Olive & Franco, 2008 for a discussion of effect sizes in single subject designs). Future studies on this topic should attempt to determine empirical guidelines for IOA across various effect sizes.

Furthermore, these data do not inform the selection of the most appropriate method of calculating IOA in clinical settings for an FA, an important consideration for clinicians and researchers. Although most of the IOA coefficients were high in the original sample and in the recalculated datasets across various cutoffs in our study, the selection of the most appropriate method of calculating IOA is complex and involves considering the type of data collected and the overall rate of responding, among other factors (Cooper et al., 2019). For example, nonoccurrence agreement may be a better measure of IOA for high rate behavior than occurrence agreement, which may be inflated with higher response rates (Bijou et al., 1968). Similarly, total agreement may be a poor measure of IOA in general because agreements are scored regardless of the actual frequency of behavior recorded by Observers A and B, provided they agreed that at least one behavior occurred (or did not occur) during an interval (Cooper et al., 2019). Thus, the total agreement method may be prone to overestimating IOA. The methods of IOA calculation used in this study were also not exhaustive, but represented some common methods that may be used by clinicians and researchers. Again, although these data may be useful in moving toward the development of empirical guidelines for determining how much IOA is enough under specific situations, these data should not be used to defend the use of inappropriate or poor measures of IOA, or the reporting of low levels of poor IOA.

These data support previous studies that have demonstrated a similar effect on the relation between response rate and IOA calculation (Mudford et al., 2009; Rolider et al., 2012). Additionally, the current results provide preliminary evidence to suggest that obtained IOA may not vary significantly with respect to the total percentage of sessions with IOA when trained observers collect data, and the target behavior is occurring at moderate-to-high rates. The overall percentage of IOA may not be as important in determining the consistency of behavioral data as other factors when comparing data collected by highly trained observers, as the obtained IOA coefficients may not be significantly different.

Although the recalculated IOA scores did not vary significantly across cutoffs regardless of IOA calculation method, it is important to consider the overall levels of variability in IOA scores for individual sessions. This effect was most significant with respect to visual inspection of the data for the occurrence agreement calculation. Original occurrence agreement was low, with significant variability in the distribution of scores for individual cases. Out of the 100 FAs included, only 46 had occurrence agreement scores above 80% and only 14 with scores above 90%. These data might support that occurrence agreement is a more stringent measure of calculating IOA (Kennedy, 2005). However, occurrence agreement is also highly sensitive to response rate in that occurrence agreement is likely to be higher when the response rate is higher (Bailey & Burch, 2018).

Given the response rates in the current sample ($M$ = 2.68 RPM, range: 0–41.16) it is possible that occurrence agreement was skewed by the variability in responding across individuals. Future studies may focus on evaluating how the IOA calculation, specifically, may be influenced by overall response rate.

Several limitations should be noted. First, the IOA cutoffs were based on existing data collected during a minimum of 30% of FA sessions by various clinical staff, so the recalculated IOA data are largely subject to the same threats that affect all data collected in clinical settings. That is, because these recalculated IOA coefficients were based on actual clinical data, they are not free from human error as might occur with fully simulated data sets. Second, although the sessions included in the recalculated IOA coefficients were randomly selected prior to analyzing the data, it is possible that a subset of sessions with high IOA were retained for analysis which influenced the findings. Third, all IOA calculations were included for all data sets; therefore, IOA calculations that were not ideal given the rate of behavior or other factors were calculated and included in subsequent analyses. Fourth, IOA was only calculated from FAs of problem behavior. It is unclear if IOA calculations would have been similar had we examined them across a more heterogeneous set of sessions. Thus, future research should examine the influence of a more heterogenous set of sessions on IOA calculations. Finally, as discussed previously, clinical data were collected by highly trained observers, in a highly structured environment, and in the specialized inpatient unit for the assessment and treatment of severe problem behavior. It is unknown if similar findings would be obtained if the original sample included clinical data from a less structured setting (e.g., schools).

One potential factor that might influence collection of reliable behavioral data is staff turnover, particularly among paraprofessional staff who may work in classrooms and homes to provide behavioral services. That is, the experience of staff may be correlated with accuracy of data collection, and efficient programs to train newly hired staff may positively impact student outcomes. For example, researchers have demonstrated success with brief behavioral skills training (BST) program to teach newly hired paraprofessional staff to correctly implement discrete trial training procedures (e.g., Catania et al., 2009; Clayton & Headley, 2019). It is plausible that these sorts of brief BST procedures, along with data collection practice, could be useful in teaching new staff to collect accurate data for common clinical procedures (e.g., functional analysis, preference assessments). Future research should examine the influence of staff expertise (e.g., students vs. experienced behavior analysts) and settings (e.g., home-based program vs. clinic) on the consistency of IOA calculations.

These data suggest that in some cases, the percentage of sessions with IOA may not be the most relevant consideration when reviewing IOA, as calculated IOA may not significantly differ—at least with clinical data collected by trained observers in a structured clinical setting. In the event similar findings are obtained using other types of clinical data from various settings, requiring fewer sessions with high IOA coefficients may encourage clinicians to include IOA, if this is not their routine practice. That is, if a perceived barrier to collecting and calculating IOA in clinical practice is that there are not enough resources to do so for "enough" sessions, these data might lead to increased interest in obtaining and subsequently calculating IOA by clinicians. Obtaining and calculating IOA for more sessions may be warranted in situations where overall IOA is variable, perhaps due to variability in the occurrence of the target behavior, or in cases where a target behavior occurs sporadically. Although reporting overall IOA coefficients of above

80% is important in establishing that our data are reliable and valid, findings from the current study suggest that reporting more sessions with IOA is not necessarily better, which may have important implications for practitioners by decreasing the burden of collecting and calculating IOA data in clinical work.

## REFERENCES


Bailey, J. S., & Burch, M. R. (2018). *Research methods in applied behavior analysis* (2nd ed.). Routledge.

Bijou, S. W., Peterson, R. F., & Ault, M. H. (1968). A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis*, *1*(2), 175–191. https://doi.org/10.1901/jaba.1968.1-175

Bullock, C. E., Fisher, W. W., & Hagopian, L. P. (2017). Description and validation of a computerized behavioral data program: "BDataPro." *The Behavior Analyst*, *40*(1), 275–285. https://doi.org/10.1007/s40614-016-0079-0

Catania, C. N., Almeida, D., Liu-Constant, B., & DiGennaro Reed, F. (2009). Video modeling to train staff to implement discrete-trial instruction. *Journal of Applied Behavior Analysis*, *42*, 387-392. http://doi.org/10.1901/jaba.2010.43-291.

Clayton, M., & Headley, A. (2019). The use of behavioral skills training to improve staff performance of discrete trial training. *Behavioral Interventions*, *34*(1), 136-143. https://doi.org/10.1002/bin.1656

Cooper, J. O., Heron, T. E., & Heward, W. L. (2019). *Applied behavior analysis* (3rd ed.). Pearson Education.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2013). *G*Power Version 1.1.7 [Computer software]*. Universität Kiel, Germany.

Harris, F. C., & Lahey, B. B. (1978). A method for combining occurrence and nonoccurrence interobserver agreement scores. *Journal of Applied Behavior Analysis*, *11*(4), 523–527. https://doi.org/10.1901/jaba.1978.11-523

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*(1), 103–116. https://doi.org/10.1901/jaba.1977.10-103

IBM Corporation (2018). IBM SPSS Statistics for Windows *(Version 25.0) [Computer software]*. IBM Corporation.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, *27*(2), 197–209. https://doi.org/10.1901/jaba.1994.27-197

Kahng, S., Ingvarsson, E. T., Quigg, A. M., Seckinger, K. E., & Teichman, H. M. (2011). Defining and measuring behavior. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), Handbook of applied behavior analysis *(pp. 113-131)*. The Guilford Press.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*(1), 141–150. https://doi.org/10.1901/jaba.1977.10-141

Kennedy, C. H. (2005). *Single-case designs for academic research*. Allyn & Bacon.

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*(1), 133–139. http://dx.doi.org/10.1901/jaba.1977.10-133

Mudford, O. C., Martin, N. R., Hui, J. K. Y., & Taylor, S. A. (2009). Assessing observer accuracy in continuous recording of rate and duration: Three algorithms compared. *Journal of Applied Behavior Analysis*, *42*(3), 527–539. https://doi.org/10.1901/jaba.2009.42-527

Olive, M. L., & Franco, J. H. (2008). (Effect) size matters: And so does the calculation. *The Behavior Analyst Today*, *9*(1), 5-10. https://doi.org/10.1037/h0100642

Page, T. J., & Iwata, B. A. (1986). Interobserver agreement: History, theory, and current methods. In A. D. Poling & R. W. Fuqua (Eds.), Research methods in applied behavior analysis: Issues and advances *(pp. 99-126)*. Plenum.

Poling, A., Methot, L., & LeSage, M. (1995). Fundamentals of behavior analytic research. New Plenum Press.

Repp, A. C., Deitz, D. E. D., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis*, *9*(1), 109–113. http://doi.org/10.1901/jaba.1976.9-109

Rolider, N. U., Iwata, B. A., & Bullock, C. E. (2012). Influences of response rate and distribution on the calculation of interobserver agreement scores. *Journal of Applied Behavior Analysis*, *45*(4), 753–762. https://doi.org/10.1901/jaba.2012.45-753

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education*, *10*(4), 205–212. http://doi.org/10.1023/A:1012295615144


*Received January 24, 2020*
*Final acceptance November 5, 2020*
*Action Editor, Jonathan Baker*