PREDICTIVE VALIDITY AND EFFICIENCY OF ONGOING VISUAL-INSPECTION CRITERIA FOR INTERPRETING FUNCTIONAL ANALYSES

Valdeep Saini, **Wayne W. Fisher**, and **Billie J. Retzlaff**

University of Nebraska Medical Center's Munroe-Meyer Institute

Abstract

Prior research has evaluated the reliability and validity of structured criteria for visually inspecting functional-analysis (FA) results on a post-hoc basis, after completion of the FA (i.e., post-hoc visual inspection [PHVI]; e.g., Hagopian et al., 1997). However, most behavior analysts inspect FAs using ongoing visual inspection (OVI) as the FA is implemented, and the validity of applying structured criteria during OVI remains unknown. In this investigation, we evaluated the predictive validity and efficiency of applying structured criteria on an ongoing basis by comparing the interim interpretations produced through OVI with (a) the final interpretations produced by PHVI, (b) the authors' post-hoc interpretations (PHAI) reported in the research studies, and (c) the consensus interpretations of these two post-hoc analyses. Ongoing visual inspection predicted the results of PHVI and the consensus interpretations with a very high degree of accuracy, and PHAI with a reasonably high degree of accuracy. Furthermore, the PHVI and PHAI results involved 32 FA sessions, on average, whereas the OVI required only 19 FA sessions to accurately identify the function(s) of destructive behavior (i.e., a 41% increase in efficiency). We discuss these findings relative to other methods designed to increase the accuracy and efficiency of FAs.

Keywords

data analysis; functional analysis; problem behavior; structured visual inspection; visual inspection

Functional analysis (FA) involves the systematic manipulation of variables hypothesized to reinforce behavior (e.g., attention, escape) using controlled, single-case experimental designs (Betz & Fisher, 2011). The development of FA methodology has allowed behavior analysts to determine the environmental variables that maintain severe destructive behavior (Beavers, Iwata, & Lerman, 2013; Hanley, Iwata, & McCord, 2003; Iwata, Dorsey, Slifer, Bauman, & Richman, 1994a;). Results of FAs have been employed to inform function-based treatments, and this process has increased the effectiveness of reinforcement-based interventions and decreased the use of punishment (Greer, Fisher, Saini, Owen, & Jones,

Address correspondence to Valdeep Saini (sainiv@upstate.edu) or Billie Retzlaff (billie. retzlaff@unmc.edu), Center for Autism Spectrum Disorders, 985450 Nebraska Medical Center, Omaha, Nebraska 68198.
Valdeep Saini is now at Upstate Medical University.

2016; Kuhn, DeLeon, Fisher, & Wilke, 1999; Pelios, Morren, Tesch, & Axelrod, 1999; Smith, Iwata, Vollmer, & Zarcone, 1992).

Researchers have noted the potential benefits of incorporating FAs into routine clinical practice (e.g., Iwata & Dozier, 2008). However, most behavior analysts continue to use indirect and descriptive assessments (Oliver, Pratt, & Normand, 2015), which tend to have poor reliability or validity compared to FAs (Kelley, LaRue, Roane, & Gadaire, 2011; Thompson & Borrero, 2011). Two of the most common reasons given by behavior analysts for using these less reliable and valid functional-assessment procedures, rather than FAs, are a lack of trained personnel and insufficient time to implement FAs (Oliver et al., 2015). As such, researchers have shown increased interest in (a) developing protocols for training clinicians to implement FAs with high procedural integrity (Iwata et al., 2000; Moore & Fisher, 2007) and (b) increasing the efficiency of FAs (e.g., Derby et al., 1997; Jessel, Hanley, & Ghaemmaghami, 2016; Kahng & Iwata, 1999).

Much of the research on training behavior analysts in FA methodology has focused on training individuals to implement various FA conditions (e.g., attention, escape) with high procedural integrity (e.g., Bloom, Lambert, Dayton, & Samaha, 2013; Iwata et al., 2000; Moore & Fisher, 2007). In most such studies, participants readily learned to implement the FA conditions with high fidelity. A second, related area of research has involved training individuals to visually inspect and accurately interpret FA results (e.g., Hagopian et al., 1997). Traditionally, visual analysis of single-subject data has been the primary method used for interpreting FA results (Betz & Fisher, 2011). Although commonly used, the reliability of visual inspection has been questioned (e.g., Fisch, 1998). Research on the reliability of visual inspection has yielded mixed results, with some studies reporting low levels of agreement between raters (DeProspero & Cohen, 1979; Danov & Symons, 2008) and others reporting high levels of agreement (Kahng, et al., 2010). One factor that may influence the reliability of visual inspection is the rater's level of experience. For example, Danov and Symons (2008) found lower average inter- and intra-rater reliability with novice raters compared to expert raters.

To address concerns related to the reliability of visual inspection during FA, as well as to aid in the interpretation of less definitive FA results, Hagopian et al. (1997) developed a set of structured visual-inspection criteria that could be used to guide interpretation of FA results. These investigators used the results obtained for the control condition of an FA (i.e., the play condition) to generate two criterion lines, one about one standard deviation (SD) above the mean and the other about one SD below the mean. They then inspected each test condition (e.g., attention) to determine how many data points fell above or below the two criterion lines. If at least five more data points fell above the upper criterion line (UCL) than below the lower criterion line (LCL), then the authors concluded that the results supported the interpretation of an operant function for the target response (e.g., concluding that an attention function was present). In addition, Hagopian et al. included additional criteria to interpret FAs with low-magnitude effects, trends in the data, or response patterns consistent with either automatic reinforcement or multiple maintaining reinforcers.

Hagopian et al. (1997) reported high agreement between interpretations based on the structured criteria and the consensus interpretation of an expert panel ($M = .94$). Additionally, predoctoral interns, who initially had low levels of agreement ($M = .54$) with the structured criteria, demonstrated high levels of agreement ($M = .90$) following training on the structured criteria. One limitation of this study was that the researchers investigated levels of visual-inspection agreement only with FA datasets with 10 data points per condition; therefore, the generality of the findings to FAs of different lengths remained unknown. To address this limitation, Roane, Fisher, Kelley, Mevers, and Bouxsein (2013) refined the structured criteria developed by Hagopian et al. to make them applicable to FA datasets of various lengths (see also Cox & Virues-Ortega, 2016).

Roane et al.'s (2013) refinements included basing decisions on the percentage of data points that fell above or below the criterion lines (rather than on an absolute number, as in Hagopian et al., 1997). In addition, Roane et al. used a function in an Excel spreadsheet to automatically generate the UCL and LCL for each dataset. They also included special rules for interpreting FAs with low-magnitude effects, trends in the data, or response patterns consistent with either automatic reinforcement or multiple maintaining reinforcers. These modifications allowed Roane et al. to apply structured, visual-inspection criteria to FA data-sets of a wide variety of lengths. The authors found that the modified criteria increased inter- and intra-rater agreement for reviewers with varying levels of expertise in applied behavior analysis (ranging from experts to postbaccalaureate employees). These findings further strengthened the conclusion that one benefit of using structured criteria for FA interpretation is an increase in agreement among raters, regardless of the rater's level of experience.

Another potential benefit of these modified visual-inspection criteria, not investigated by Roane et al. (2013), is that the criteria might be used to guide ongoing visual inspection (OVI) of FA results in real time as the FA is conducted. That is, behavior analysts could apply structured visual-inspection criteria after each repetition of a set of test and control conditions or even after every session (cf. Ferron, Joo, & Levin, 2017; Vanselow, Thompson, & Karsina, 2011). For example, Figure 1 is a reprint of the graph in the bottom panel of Figure 1 from Hagopian, Toole, Long, Bowman, and Lieving (2004). It shows 10 sessions each for four test conditions and one control condition (50 sessions total). Application of the Roane et al. criteria on an ongoing basis after each session, would yield the same conclusion (i.e., presence of an escape function) after 28 sessions (i.e., a 44% increase in efficiency). To date, no studies have investigated the predictive validity of implementing structured criteria on an ongoing basis relative to interpretations provided after the FA has concluded.

In most situations, the results of a new assessment or diagnostic test are compared with those obtained from the best available assessment, and the level of agreement between the new assessment and the best assessment is used to determine the predictive or concurrent validity of the new assessment (Akobeng, 2007; Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008). Some behavior analysts might consider structured visual-inspection criteria applied on a post-hoc basis to be the best criterion for interpreting FA results, given the demonstrated agreement between this approach and the interpretations of an expert panel. On the other hand, other behavior analysts might consider post-hoc interpretations provided by the authors of FA studies (post-hoc author interpretation [PHAI]) the best criterion measure

because these interpretations have been subjected to the peer-review process associated with publication of the investigations. We therefore elected to compare interpretations derived from OVI with both PHVI and PHAI, and also with the consensus agreement between these two post-hoc visual inspection methods for those FAs on which PHVI and PHAI agreed.

In summary, the primary purpose of the current study was to evaluate the predictive validity of OVI using the Roane et al. (2013) criteria in comparison to three criterion measures (PHVI, PHAI, and the consensus interpretations of PHVI and PHAI for those FAs on which they agreed). The secondary purpose of this study was to determine whether OVI using the Roane et al. criteria would increase the efficiency of FAs (i.e., allow behavior analysts to reach conclusions comparable to these criterion measures in a fewer number of sessions and a lesser amount of time).

METHOD

Article Identification

We included all studies reviewed by Hanley et al. (2003) and Beavers et al. (2013; citations obtained from study authors). We then conducted a search of the published literature from 2011 through 2015 using search criteria based on those described by Hanley et al. We searched PsycINFO, ERIC, the *Journal of Applied Behavior Analysis*, and *Google Scholar*, using the keywords *functional analysis*, *functional assessment*, and *behavioral assessment*. We also examined the reference section of each obtained article to identify additional studies not identified during the initial search.

Inclusion and Exclusion Criteria for Dataset Identification

We used random sampling to identify data-sets to be included in our analysis from the articles described above. We inserted the names of these studies into a list randomizer in alphabetical order beginning with the last name of the first author. We randomized the list prior to each evaluation and evaluated only the datasets within the study that appeared at the top of the randomized list. Next, we selected one or more datasets from the first study on the randomized list if the dataset(s) met the following criteria: (a) the FA used reliable, direct-observation measures (i.e., interobserver agreement 80%) and a multielement design to compare the conditions of the FA; (b) the FA included at least six sessions per condition from at least three different conditions (i.e., minimum of 18 sessions), with one of the conditions being a control condition (e.g., play); and (c) the FA results were depicted in graphical form showing individual data points from each session. We excluded data from supplemental analyses that did not meet the above criteria (e.g., pairwise comparison; Iwata, Duncan, Zarcone, Lerman, & Shore, 1994b).

After evaluating the FAs from the first study on the randomized list, we re-randomized the remaining articles on the list (i.e., randomization without replacement) and again evaluated the FAs in the first article on the re-randomized list. We repeated this process until we had selected 100 FAs that met the above inclusion criteria (see supplemental material for full list of FAs included). Figure 2 provides exemplary cases of extended analyses with at least six

sessions per condition (reprinted from Hanley, Piazza, Fisher, & Maglieri, 2005) that were evaluated in the present study.

Data Extraction

We used PlotDigitizer (Huwaldt, 2016) to extract data from imported images of the graphs in PDF format. First, we identified the *x* and *y* coordinates and scale of the axes by clicking and entering values for the origin, the maximum value of the *x*-axis, and the maximum value of the *y*-axis. We then extracted the value of every data point of every session in the analysis. To obtain the value of each datum, we clicked on the center of each datum, which produced corresponding *x* and *y* coordinates. Given that we first established the axes dimensions for each dataset, this method proved applicable for all types of measures (e.g., response rate, percentage of intervals). We extracted the data values produced by PlotDigitizer into a Microsoft Excel spreadsheet. Finally, for each numerical dataset within the Excel spreadsheet, we created a standardized graph that included the UCL and LCL.

Procedure

We evaluated all 100 standardized graphs using the Roane et al. (2013) criteria both on an ongoing basis (OVI) and on a post-hoc basis (PHVI), including the application of the special rules for FAs with low-magnitude effects, trends in the data, or response patterns consistent with either automatic reinforcement or multiple maintaining reinforcers. We also tabulated the post-hoc interpretations provided by the authors of each article (PHAI).

Post-hoc visual inspection—Using the formulae and guides described in the Appendix of Roane et al. (2013), we developed a Microsoft Excel workbook that automatically inserted the UCL and LCL into the graph. The experimenter then applied the Roane et al. criteria for each potential function in each dataset (e.g., if the FA dataset tested for an attention and an escape function, the experimenter applied the Roane et al. [2013] criteria to each of those two potential functions). That is, for each potential function, the experimenter (a) counted the number of data points that fell above the UCL (*u*); (b) counted the number of data points that fell below the LCL (*l*); (c) counted the number of sessions conducted to test that function (*n*); and (d) used the values for *u*, *l*, and *n* to determine whether the following equation was true:

$$u - l \geq n/2$$

When the above equation held true for a given potential function (e.g., escape) in a given dataset, the experimenter concluded that the tested consequence functioned as a reinforcer for destructive behavior (e.g., destructive behavior was reinforced by escape). For example, suppose that an FA dataset included seven escape sessions, six attention sessions, and six control sessions (19 sessions total). Suppose further that five escape sessions fell above the UCL and one escape sessions fell below the LCL. When the experimenter entered these values into the above formula, the formula held true (5 − 1   7/2), and we concluded that an escape function was present. Finally, suppose that three attention sessions fell above the UCL and two attention sessions fell below the LCL. When the experimenter entered these values into the above formula, the formula did not hold true (3 − 2 is not   6/2), and we

concluded that an attention function was not present. After applying the above formula, we also evaluated the FA relative to the special rules for FAs with low-magnitude effects, trends in the data, or response patterns consistent with either automatic reinforcement or multiple maintaining reinforcers (see appendix in Roane et al., 2013 for details).

On-going visual inspection—The procedures used for OVI paralleled those for PHVI with the following exceptions. Initially, the experimenter applied the Roane et al. (2013) criteria to only the first three data points ($n = 3$) from each condition. For example, suppose for the escape condition described above that one of the first three data points fell above the UCL ($u = 1$) and zero fell below the LCL ($l = 0$). In this scenario, the above equation would not hold true ($1 - 0$ is not $\geq 3/2$). If the equation did not hold true for any of the potential functions with three data point per condition, we repeated this process with the first four data points ($n = 4$) from each condition. For example, suppose for the escape condition described above that three of the first four data points fell above the UCL ($u = 3$) and one fell below the LCL ($l = 1$). In this scenario, the above equation would hold true ($3 - 1 \geq 4/2$).

Once we identified at least one function for which the formula held, we evaluated whether we should discontinue OVI. We discontinued OVI if (a) we identified at least one function for destructive behavior (e.g., escape; escape and tangible), and (b) we would not identify an additional function by extending the FA by one data point per condition (e.g., $n = 5$ data points per condition). For example, suppose that we first identified only an escape function after evaluating an FA with four data points per condition (i.e., $n = 4$). We would then evaluate whether the equation would hold true for each other function (e.g., attention) if the next data point for that function fell above the UCL. Suppose further that two attention sessions fell above the UCL and one fell below the LCL when we applied the OVI criteria with four sessions per condition. Given these values, even if the next attention condition fell above the UCL, we would not identify an attention function (i.e., if $u = 3$, $l = 1$ and $n = 5$, then $3 - 1$ is not $\geq 5/2$). We conducted this "what if" analysis for each potential function, and we discontinued OVI if it was not possible for us to identify an additional function by increasing the length of OVI by one session per condition.

Data Analysis

We used levels of exact agreement as the primary dependent variable to compare the accuracy of OVI with three criterion measures: (a) PHVI, (b) PHAI, and (c) the consensus agreement between PHVI and PHAI for those FAs that these two post-hoc methods agreed. That is, we determined how often OVI produced the same interpretation as each criterion measure for each FA. We also evaluated correspondence between OVI and these three criterion measures using dependent measures commonly used by other disciplines to evaluate the predictive validity of a new assessment or diagnostic test. That is, we evaluated the validity of OVI using four commonly used statistical measures: (a) sensitivity, (b) specificity, (c) positive predictive value (PPV), and (d) negative predictive value (NPV; Akobeng, 2007). In the case of functional analysis, sensitivity refers to the proportion of individuals who show a particular function of behavior (e.g., attention-reinforced destructive behavior) that are identified as showing that function by a new assessment procedure such as OVI. Specificity refers to the proportion of individuals who do *not* show a function (e.g.,

absence of escapereinforced destructive behavior) that are identified as *not* showing that function by the new assessment. The PPV statistic represents the probability that an individual identified as showing a particular function actually does. The NPV statistic represents the probability that an individual identified as not showing a particular function actually does not.

Interrater Agreement

Two independent raters evaluated 34% of the datasets that met inclusion criteria. Each rater applied the PHVI and OVI criteria to a given dataset, and we assessed reliability by comparing results of each rater's PHVI and OVI analysis (i.e., we identified whether the two raters arrived at the same conclusion within the same number of analyzed sessions). We calculated interrater agreement by dividing the number of agreements by the number of agreements plus disagreements and converting the resulting proportion to a percentage for each dataset, resulting in a mean interrater agreement of 97% across datasets.

One dataset met criteria for destructive behavior maintained by automatic and multiple types of social reinforcement, which produced the single disagreement among raters. To resolve this disagreement, a third, independent rater analyzed the dataset and that rater's interpretation (i.e., automatic reinforcement) was the one used during our evaluation.

Measurement of FA Efficiency

In addition to recording the identified function(s) of destructive behavior using the various visual-inspection methods, we also quantified and compared these methods relative to the number of sessions and amount of time required to complete each type of visual inspection. For the measure of assessment duration, we multiplied the number of sessions analyzed by the duration of each session as described by the authors of the respective studies.

RESULTS

Application of the inclusion criteria resulted in the identification of a random sample of 100 multielement FAs from 57 studies published between 1990 and 2014. Within the 100 FAs, we evaluated levels of exact agreement between OVI and the three criterion measures as well as levels of agreement and disagreement on each of the 321 individual operant functions.

Exact Agreement

We defined an exact agreement as an FA in which multiple interpretation methods (e.g., OVI and PHVI) identified the same set of functions and nonfunctions (e.g., both OVI and PHVI identified the presence of an attention and an escape function and the absence of an automatic function). We calculated the percentage of exact agreements by dividing the number of FAs with correspondence (e.g., all functions and nonfunctions identified by PHVI also identified by OVI) by the total number of FAs analyzed (i.e., 100).

Table 1 shows the number and percentage of exact agreements by function, between OVI in comparison with the three criterion measures: (a) PHVI, (b) PHAI, and (c) the consensus of

PHVI and PHAI. It should be noted that when we used the consensus interpretations of PHVI and PHAI as the criterion measure, we included only the 83 FAs on which PHVI and PHAI produced the same interpretation.

The top panel of Table 1 shows that OVI achieved high levels of exact agreement with PHVI, producing the same interpretation for 93 of the 100 FAs. Most disagreements (i.e., 4 of 7) consisted of FAs in which PHVI identified multiple functions and OVI identified only one function or identified a different set of multiple functions.

As can be seen in the middle panel of Table 1, OVI showed moderately high levels of exact agreement with PHAI, producing the same interpretation for 80 of 100 FAs. Approximately one half of the disagreements (i.e., 11 of 20) consisted of FAs in which PHAI identified a single social reinforcement function (i.e., attention, tangible, or escape) and OVI identified multiple functions. In many of these cases, the authors provided an incomplete interpretation of the functional analysis, stating that the participant's behavior was sensitive to a specific form of social reinforcement without commenting directly on other possible functions. Figure 3 shows an example of an FA in which the authors identified an attention function, but did not comment on other possible functions (data reprinted from Fisher, Ninness, Piazza, & Owen-DeSchryver, 1996). In this case, both OVI and PHVI identified one additional function (i.e., tangible).

The bottom panel of Table 1 shows that OVI achieved high levels of exact agreement with the consensus agreements between PHVI and PHAI. Of the 83 FAs for which PHVI and PHAI showed exact agreement, OVI produced the same agreement for 79 (95.2%) of these FAs.

Sensitivity, Specificity, and Positive and Negative Predictive Values

Table 2 shows the number of times the OVI agreed and disagreed with each criterion on the presence and absence of individual functions of destructive behavior. We used the following measures to determine the predictive validity of OVI for identifying individual functions with the three criterion measures: (a) PHVI, (b) PHAI, and (c) consensus agreement between PHVI and PHAI. We used the numbers presented in Table 2 to calculate the following measures.

Sensitivity—Sensitivity is the proportion (or percentage) of all functions identified by the criterion measures also identified by OVI. Using values from Table 2, sensitivity = a / (a + c).

Specificity—Specificity is the proportion of all nonfunctions identified by the criterion measures also identified by OVI. Using values from Table 2, specificity = d / (b + d).

Positive predictive value—Positive predictive value is the proportion of functions identified by OVI also identified by the criterion measures. Using values from Table 2, PPV = a / (a + b).

Negative predictive value—Negative predictive value is the proportion of nonfunctions identified by OVI also identified by the criterion measures. Using values from Table 2, NPV = d / (c + d).

Total agreement—Total agreement is the proportion of functions tested in which the results of the OVI agreed with the results of the criterion measure regardless of whether the criterion measure identified the presence or absence of a function. Using the values from Table 2, total agreement = (a + d) / (a + b + c + d).

Figure 4 displays the percentage of agreement between OVI and each of the three criterion measures on commonly used measures of predictive validity. The black bars indicate the percentages of agreement between OVI and PHVI on these measures of predictive validity. Relative to PHVI, OVI produced high levels of sensitivity (96.3%), specificity (97.3%), PPV (96.3%,), and NPV (97.3%), with a total agreement of 96.9%. Relative to PHVI, OVI showed no bias toward identifying a function not identified by PHVI ($n = 5$) versus failing to identify a function identified by PHVI ($n = 5$; see the top panel of Table 2).

The grey bars in Figure 4 depict the percentage of agreement between OVI and PHAI. Relative to PHAI, OVI showed slightly lower, but still high levels of sensitivity (86.0%), specificity (95.7%), PPV (93.6%), NPV (90.3%), and total agreement (91.6%). Relative to PHAI, OVI showed a bias toward failure to identify a function identified by PHAI (n = 19) versus identifying a function not identified by PHAI (n = 8; see the middle panel of Table 2).

The white bars in Figure 4 indicate the percentage of agreement between OVI and the consensus agreement between PHVI and PHAI. Relative to the consensus, OVI showed high levels of sensitivity (99.0%), specificity (96.9%), PPV (95.4%), NPV (99.4%), and total agreement (97.7%). Relative to the consensus, OVI showed a slight bias toward failing to identify a function identified by the consensus (n = 5) versus identifying a function not identified by the consensus (n = 1; see the bottom panel of Table 2), although the total number of disagreements with the consensus was quite small (n = 6).

Quantitative and Qualitative Analysis of Exact Agreements and Disagreements

Table 3 shows the number of FAs for which: (a) OVI, PHVI, and PHAI all produced the same interpretations; (b) two interpretive methods produced the same interpretation (e.g., PHVI and PHAI), but the other disagreed (e.g., OVI); and (c) all three interpretive methods produced different interpretations. OVI, PHVI, and PHAI all agreed on 79 of the 100 FAs, and these three interpretive methods each produced distinctly different interpretations on just two of the 100 FAs. Figure 5 shows an example of an FA in which all three methods produced the same interpretation in the top panel (data reprinted from Piazza et al., 1999) and an example of an FA in which all three interpretive methods produced uniquely different interpretations in the bottom panel (data reprinted from Falcomata, Muething, Gainey, Hoffman, & Fragale, 2013). As can be seen in the top panel of Figure 5, the interpretation of the FA with consensus agreement across OVI, PHVI, and PHAI is relatively simple and straightforward. By contrast, the FA in the bottom panel show the highest levels of destructive behavior during the demand condition in the first half of the analysis and the highest levels in the tangible condition during the second half of the analysis. The PHAI

identified both an escape and a tangible function. The OVI met the criteria for discontinuation of the analysis after 14 sessions (i.e., about three sessions in each condition), and thus identified only an escape function. The PHVI identified a tangible function based on the special rules for trend because the tangible data path showed an upward trend in the last half of the analysis but did not identify a escape function (4–2 is not ≥ 6/2).

Table 3 also shows that when two of the three interpretive methods produced exact agreements, those two methods were most commonly OVI and PHVI, with PHAI providing a disparate interpretation ($n = 14$). In many of these cases ($n = 8$ of 14; 57.1%), the authors identified one function or multiple functions without commenting on the presence or absence of at least one other potential function, as indicated earlier and illustrated in Figure 3.

Figure 6 shows an example of an FA dataset for which PHVI and PHAI showed exact agreement but OVI resulted in a different interpretation (data reprinted from Hanley, Piazza, & Fisher, 1997). Early in the FA, the participant displayed destructive behavior mostly in the demand condition. OVI identified an escape function and met criteria for discontinuation after 26 sessions. However, as can be seen, rates of destructive behavior subsequently decreased to low levels in the demand condition and increased markedly in the attention condition, resulting in both post-hoc interpretations indicating an attention function for destructive behavior.

Figure 7 shows an example of an FA dataset for which PHVI and OVI showed exact agreement but PHAI resulted in a different interpretation (data reprinted from Smith, Iwata, Goh, & Shore, 1995). The authors interpreted (i.e., PHAI) the results as indicating an escape function for this participant's SIB, but both OVI and PHVI interpreted the results as undifferentiated, due to control sessions with unusually high levels of SIB. The authors hypothesized that the high levels of SIB in the control condition probably resulted from carry-over effects due to similarities between the demand and control conditions (i.e., multiple treatment interference). Strict application of the structured criteria, either on an ongoing or post-hoc bases, does not factor in such qualitative interpretations.

Figure 8 shows an example of an FA dataset for which PHAI and OVI showed exact agreement but PHVI resulted in a different interpretation (data re-printed from Adelinis, Piazza, Fisher, & Hanley, 1997). Using PHVI, we identified the FA as undifferentiated due to the variability in the latter half of the analysis. Using OVI, we identified an attention function early on and met the criteria for discontinuation of analysis after 24 sessions, before the variability became apparent. Using PHAI, the authors indicated that the results suggested an attention function, and they subsequently conducted additional analyses that showed that the presence of an attention function became more apparent in the presence of particular establishing operations. That is, the interpretation of an attention function appeared to be correct and the presence or absence of relevant establishing operations accounted for the initial variability.

Functional Analyses That Did Not Meet Termination Criteria

For 13 datasets, we repeatedly applied the OVI. No operant function was identified with three sessions per condition; we lengthened and reanalyzed the results until we analyzed the full-length FA and produced the same outcome as PHVI. In four of these datasets, neither OVI nor PHVI identified a function for the target destructive behavior (see Figure 7 for a representative graph).

Efficiency of OVI compared to Post-Hoc Analyses

Across datasets, OVI required a mean of 19 sessions (range, 9 to 83) compared to a mean of 32 sessions (range, 18 to 83) for the post-hoc analyses. The average assessment duration when using OVI was 198 min (range, 45 to 1245 min), compared to a mean of 326 min (range, 75 to 1245 min) for the post-hoc analyses. This represent a 41% reduction in the total number of sessions required to accurately identify functions of destructive behavior and a 39% reduction in total time required to conduct the assessment. If the authors of the sampled studies had used OVI as an aid during FA, the average FA duration could have been decreased by 128 min.

DISCUSSION

We evaluated the accuracy and efficiency of structured visual-inspection criteria applied on an ongoing basis compared to three post-hoc criterion measures: (a) structured criteria applied on a post-hoc basis (i.e., PHVI), (b) the authors' interpretations applied on a post-hoc basis (i.e., PHAI), and (c) the consensus agreement between PHVI and PHAI (for those FAs on which they agreed). OVI produced very high levels of exact agreement with PHVI and with the consensus between PHVI and PHAI, and moderately high levels of exact agreement with PHAI (93% with PHVI, 95.2% with the consensus, 80% with PHAI). When considering the total agreement on the presence or absence of the 321 functions evaluated, OVI showed even higher levels of agreement (96.9% with PHVI, 97.7% with the consensus, 91.6% with PHAI). Overall, these results suggest that OVI represents an accurate method of interpreting FA results as the FA is conducted.

It is worth noting that we observed a larger discrepancy between the total-agreement measure and the exact-agreement measure when we used PHAI as the criterion measure (i.e., an 11.6% difference) than when we used PHVI as the criterion measure (i.e., a 3.9% difference). In fact, across all measures of agreement between OVI and the criterion measures, we observed higher levels of correspondence between OVI and PHVI than between OVI and PHAI and between PHVI and PHAI. For example, amongst the 21 FAs for which we observed at least one disagreement, the majority of those (14; 66.7%) involved exact agreements between OVI and PHVI but disagreements with PHAI. These results are consistent with the findings of prior studies showing that the use of structured criteria can improve the reliability of visual inspection (e.g., Hagopian et al., 1997; Fisher, Kelley, & Lomas., 2003; Roane et al., 2013).

The current results are also noteworthy in that we included measures commonly used by other disciplines to evaluate the predictive validity of a new assessment or diagnostic test.

That is, we evaluated the validity of OVI using four commonly used statistical measures: (a) sensitivity, (b) specificity, (c) PPV, and (d) NPV (Akobeng, 2007).

When sensitivity is high, as we observed for OVI, it suggests that we are identifying almost all of the individuals whose destructive behavior is actually maintained by the reinforcer tested, but we also may be obtaining positive results with a substantial number of individuals whose problem behavior is not maintained by the reinforcer tested (i.e., false positives). Therefore, when sensitivity is high (e.g., 96.3% for OVI with PHVI as the criterion measure), we can have confidence when the new assessment produces a negative result (e.g., no escape function), because very few of the individuals whose problem behavior is not maintained by the reinforcer tested produce a positive result (i.e., 3.7%). However, when the new assessment produces a positive result, we should have less confidence in the results based solely on a high sensitivity value. By contrast, when specificity is high (e.g., 97.3% for OVI with PHVI as the criterion measure), we should have confidence when the new assessment produces a positive result (e.g., presence of an escape function), because very few of the individuals whose destructive behavior tests negative for a given function actually have destructive behavior maintained by that particular reinforcer (i.e., 2.7%).

The sensitivity and specificity statistics are useful primarily for evaluating the overall accuracy of a new assessment with a group of individuals, but they are much less useful when evaluating the accuracy of a result obtained from the new assessment for an individual patient (Akobeng, 2007). That is, a clinician typically wants to know how often the result is correct and how often it is incorrect when evaluating a positive (or negative) result obtained with an individual patient; the PPV and NPV statistics are useful for such purposes. Our OVI showed high PPV and high NPV, with all values above 90% for all criterion measures. Across all of the measures used to evaluate predictive validity (sensitivity, specificity, PPV, NPV, total agreement, exact agreement), OVI performed well with all criterion measures, but particularly well when either PHVI or the consensus between PHVI and PHAI served as the criterion measure.

The current results also are potentially important because OVI is more consistent with how behavior-analysis practitioners interpret FA results. That is, practitioners visually inspect FA results on an ongoing basis as the FA is conducted and cease the FA when they obtain relatively clear and interpretable results (see also Vanselow et al., 2011). The current results suggest that by using OVI, practitioners may reach this endpoint (i.e., relatively clear and interpretable FA results) more efficiently, and they can have reasonable confidence that the obtained results will match the results of a more extended FA. However, it is possible that the authors of the articles from which the included FA datasets originated extended their FAs for reasons beyond simply identifying the maintaining variables of destructive behavior in an efficient manner (e.g., to show an effect or to use the extended FA as a point of comparison). One limitation of the present study was that we only included temporally extended Fas (i.e., those with at least six sessions per condition from at least three different conditions). This inclusion criterion may have artificially inflated the efficiency of OVI as we necessarily compared OVI to longer FAs (i.e., those with a greater number of sessions). Using extended FAs was necessary for us to reliably apply OVI and to compare the outcome of OVI to PHVI.

Recently, investigators have shown increased interest in improving the efficiency of various FA methods (e.g., Jessel et al., 2016). Our OVI method proved to be efficient without substantially compromising accuracy. For example, when comparing the effectiveness of OVI and PHVI at predicting the authors' interpretations (i.e., PHAI), OVI increased efficiency by over 40% relative to PHVI and obtained exact agreements with PHAI for 80% of the FAs, compared to 83% for PHVI. Thus, OVI produced a relatively large increase in efficiency with only a 3% decrement in accuracy.

Investigators have also shown increased interest in comparing various FA methods in terms of efficiency (e.g., Jessel et al., 2016). The OVI methods evaluated in the current study may be useful for such comparisons. That is, OVI provides objective criteria for determining the minimum number of FA sessions required to produce relatively clear and interpretable results. Thus, OVI could potentially be used to objectively compare two different FA methods to determine which one produced reasonably clear and interpretable results in fewer sessions.

Increasing the efficiency of FAs using OVI may also have practical implications. That is, practicing behavior analysts do not often conduct FAs because of the time required to complete such assessments (Oliver et al., 2015). Ongoing visual inspection could be used as an interpretive aid to assist in potentially identifying functions of destructive behavior sooner than might be the case without OVI. However, it should be noted that OVI should be used only in an assistive capacity and not in lieu of best clinical judgement or rigorous experimental evaluation. Further, we do not recommend behavior analysts use OVI as a substitute for appropriate training in visual inspection and single-subject data analysis. In accordance with the recommendations made by Hagopian et al. (1997) and Roane et al. (2013), the use of structured visual-inspection criteria should not be used as the only method for interpreting FA data.

Although the results of the present study show promise for the utility of OVI during FA, its utility may be limited in some cases. For example, most disagreements between OVI and the three criterion measures resulted from emerging or diminishing behavioral trends that became apparent only during the extended assessment. In this respect, OVI may fail to identify a function of destructive behavior that emerges only after repeated exposure to FA conditions, or it may falsely identify a function of destructive behavior that appears to occur when the assessment is initiated and subsequently decreases after repeated exposure to FA conditions. However, these disagreements between OVI and post-hoc analyses may have been a product of how OVI was applied in the present study. That is, we terminated OVI when it first identified one or more functions of destructive behavior and when conducting one more block of sessions would not have identified additional functions. Had we continued to apply OVI based on the presence of potential emerging trends in the data, we may have seen even greater correspondence between OVI and post-hoc analyses. Future research should determine whether additional structured criteria could be developed to accurately identify emerging trends in an FA dataset during OVI and to adjust the length of the FA accordingly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Adelinis JD, Piazza CC, Fisher WW, Hanley GP. The establishing effects of client location on self-injurious behavior. Research in Developmental Disabilities. 1997; 18:383–391. https://doi.org/10.1016/S0891-4222(97)00017-6. [PubMed: 9292931]

Akobeng AK. Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. Acta Paediatrica. 2007; 96:338–341. https://doi.org/10.1111/j.1651-2227.2006.00180.x. [PubMed: 17407452]

Beavers GA, Iwata BA, Lerman DC. Thirty years of research on the functional analysis of problem behavior. Journal of Applied Behavior Analysis. 2013; 46:1–21. https://doi.org/10.1002/jaba.30. [PubMed: 24114081]

Betz, AM., Fisher, WW. Functional analysis: History and methods. In: Fisher, WW.Piazza, CC., Roane, HS., editors. Handbook of Applied Behavior Analysis. New York, NY: Guilford; 2011. p. 206-225.

Bloom SE, Lambert JM, Dayton E, Samaha AL. Teacher-conducted trial-based functional analyses as the basis for intervention. Journal of Applied Behavior Analysis. 2013; 46:208–218. https://doi.org/10.1002/jaba.21. [PubMed: 24114095]

Cox AD, Virues-Ortega J. Interactions between behavior function and psychotropic medication. Journal of Applied Behavior Analysis. 2016; 49:85–104. https://doi.org/10.1002/jaba.247. [PubMed: 26391410]

Danov SE, Symons FJ. A survey evaluation of the reliability of visual inspection and functional analysis graphs. Behavior Modification. 2008; 32:828–839. https://doi.org/10.1177/0145445508318606. [PubMed: 18490266]

DeProspero A, Cohen S. Inconsistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis. 1979; 12:573–579. https://doi.org/10.1901/jaba.1979.12-573. [PubMed: 16795617]

Derby KM, Wacker DP, Berg W, DeRaad A, Ulrich S, Asmus J, Stoner EA. The long-term effects of functional communication training in home settings. Journal of Applied Behavior Analysis. 1997; 30:507–531. https://doi.org/10.1901/jaba.1997.30-507. [PubMed: 9316262]

Falcomata TS, Muething CS, Gainey S, Hoffman K, Fragale C. Further evaluation of functional communication training and chained schedules of reinforcement to treat multiple functions of challenging behavior. Behavior Modification. 2013; 37:723–746. https://doi.org/10.1177/0145445513500785. [PubMed: 23990645]

Ferron JM, Loo SH, Levin JR. A Monte-Carlo evaluation of masked-visual analysis in responseguided versus fixed-criteria multiple-baseline designs. Journal of Applied Behavior Analysis. 2017; 50:701–716. https://doi.org/10.1002/jaba.410. [PubMed: 28887866]

Fisch GS. Visual inspection of data revisited: Do the eyes still have it? The Behavior Analyst. 1998; 21:111–123. [PubMed: 22478303]

Fisher WW, Kelley ME, Lomas JE. Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. Journal of Applied Behavior Analysis. 2003; 36:387–406. [PubMed: 14596583]

Fisher WW, Ninness HAC, Piazza CC, Owen-DeSchryver JS. On the reinforcing effects of the content of verbal attention. Journal of Applied Behavior Analysis. 1996; 29:235–238. https://doi.org/10.1901/jaba.1996.29-235. [PubMed: 8682738]

Greer BD, Fisher WW, Saini V, Owen TM, Jones JK. Improving functional communication training during reinforcement schedule thinning: An analysis of 25 applications. Journal of Applied Behavior Analysis. 2016; 49:105–121. https://doi.org/10.1002/jaba.265. [PubMed: 26482103]

Hanley GP, Iwata BA, McCord BE. Functional analysis of problem behavior: A review. Journal of Applied Behavior Analysis. 2003; 36:147–185. https://doi.org/10.1901/jaba.2003.36-147. [PubMed: 12858983]

Hanley GP, Piazza CC, Fisher WW. Noncontingent presentation of attention and alternative stimuli in the treatment of attention-maintained destructive behavior. Journal of Applied Behavior Analysis. 1997; 30:229–237. https://doi.org/10.1901/jaba.1997.30-229. [PubMed: 9210303]

Hanley GP, Piazza CC, Fisher WW, Maglieri KA. On the effectiveness of and preference for punishment and extinction components of function-based interventions. Journal of Applied Behavior Analysis. 2005; 38:51–65. https://doi.org/10.1901/jaba.2005.6-04. [PubMed: 15898474]

Hagopian LP, Fisher WW, Thompson RH, Owen-DeSchryver J, Iwata BA, Wacker DP. Toward the development of structured criteria for interpretation of functional analysis data. Journal of Applied Behavior Analysis. 1997; 30:313–326. https://doi.org/10.1901/jaba.1997.30-313. [PubMed: 9210309]

Hagopian LP, Toole LM, Long ES, Bowman LG, Lieving GA. A comparison of dense-to-lean and fixed lean schedules of alternative reinforcement and extinction. Journal of Applied Behavior Analysis. 2004; 37:323–338. https://doi.org/10.1901/jaba.2004.37-323. [PubMed: 15529889]

Huwaldt, JA. Plot Digitizer. 2016. Retrieved from https://sourceforge.net/projects/plotdigitizer/

Iwata BA, Dorsey MF, Slifer KJ, Bauman KE, Richman GS. Toward a functional analysis of self-injury. Journal of Applied Behavior Analysis. 1994a; 27:197–209. https://doi.org/10.1901/jaba.1994.27-197. [PubMed: 8063622]

Iwata BA, Dozier CL. Clinical application of functional analysis methodology. Behavior Analysis in Practice. 2008; 1:3–9. [PubMed: 22477673]

Iwata BA, Duncan BA, Zarcone JR, Lerman DC, Shore BA. A sequential, test-control methodology for conducting functional analyses of self-injurious behavior. Behavior Modification. 1994b; 18:289–306. https://doi.org/10.1177/01454455940183003. [PubMed: 8037650]

Iwata BA, Wallace MD, Kahng S, Lindberg JS, Roscoe EM, Conners J, Worsdell AS. Skill acquisition in the implementation of functional analysis methodology. Journal of Applied Behavior Analysis. 2000; 33:181–194. https://doi.org/10.1901/jaba.2000.33-181. [PubMed: 10885526]

Jessel J, Hanley GP, Ghaemmaghami M. Interview-informed synthesized contingency analyses: Thirty replications and reanalysis. Journal of Applied Behavior Analysis. 2016; 49:576–595. https://doi.org/10.1002/jaba.316. [PubMed: 27174653]

Kahng S, Chung K, Gutshall K, Pitts SC, Kao J, Girolami K. Consistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis. 2010; 43:35–45. https://doi.org/10.1901/jaba.2010.43-35. [PubMed: 20808494]

Kahng S, Iwata BA. Correspondence between outcomes of brief and extended functional analyses. Journal of Applied Behavior Analysis. 1999; 32:149–159. https://doi.org/10.1901/jaba.1999.32-149. [PubMed: 10396768]

Kelley, ME., LaRue, R., Roane, HS., Gadaire, DM. Indirect behavioral assessments: Interviews and rating scales. In: Fisher, WW.Piazza, CC., Roane, HS., editors. Handbook of Applied Behavior Analysis. New York, NY: Guilford; 2011. p. 182-190.

Kuhn DE, DeLeon IG, Fisher WW, Wilke AE. Clarifying an ambiguous functional analysis with matched and mismatched extinction procedures. Journal of Applied Behavior Analysis. 1999; 32:99–102. https://doi.org/10.1901/jaba.2010.43-35. [PubMed: 10201106]

Moore JW, Fisher WW. The effects of videotape modeling on staff acquisition of functional analysis methodology. Journal of Applied Behavior Analysis. 2007; 40:197–202. https://doi.org/10.1901/jaba.2007.24-06. [PubMed: 17471805]

Oliver AC, Pratt LA, Normand MP. A survey of functional behavior assessment methods used by behavior analysts in practice. Journal of Applied Behavior Analysis. 2015; 48:817–829. https://doi.org/10.1002/jaba.256. [PubMed: 26411336]

Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology. 2008; 56:45–50. https://doi.org/10.4103/0301-4738.37595. [PubMed: 18158403]

Pelios L, Morren J, Tesch D, Axelrod S. The impact of functional analysis methodology on treatment choice for self-injurious and aggressive behavior. Journal of Applied Behavior Analysis. 1999; 32:185–195. https://doi.org/10.1901/jaba.1999.32-185. [PubMed: 10396771]

Piazza CC, Bowman LG, Contrucci SA, Delia MD, Adelinis JD, Goh H. An evaluation of the properties of attention as reinforcement for destructive and appropriate behavior. Journal of Applied Behavior Analysis. 1999; 32:437–449. https://doi.org/10.1901/jaba.1999.32-437. [PubMed: 10641299]

Roane HS, Fisher WW, Kelley ME, Mevers JL, Bouxsein KJ. Using modified visual-inspection criteria to interpret functional analysis outcomes. Journal of Applied Behavior Analysis. 2013; 46:130–146. https://doi.org/10.1002/jaba.13. [PubMed: 24114090]

Smith RG, Iwata BA, Goh H, Shore BA. Analysis of establishing operations for self-injury maintained by escape. Journal of Applied Behavior Analysis. 1995; 28:515–535. https://doi.org/10.1901/jaba.1995.28-515. [PubMed: 16795880]

Smith RG, Iwata BA, Vollmer TR, Zarcone JR. Experimental analysis and treatment of multiply controlled self-injury. Journal of Applied Behavior Analysis. 1992; 26:183–196. https://doi.org/10.1901/jaba.1993.26-183.

Thompson, RH., Borrero, JC. Direct observation. In: Fisher, WW.Piazza, CC., Roane, HS., editors. Handbook of Applied Behavior Analysis. New York, NY: Guilford; 2011. p. 191-205.

Vanselow NR, Thompson R, Karsina A. Data-based decision making: The impact of data variability, training, and context. Journal of Applied Behavior Analysis. 2011; 44:767–780. https://doi.org/10.1901/jaba.2011.44-767. [PubMed: 22219528]
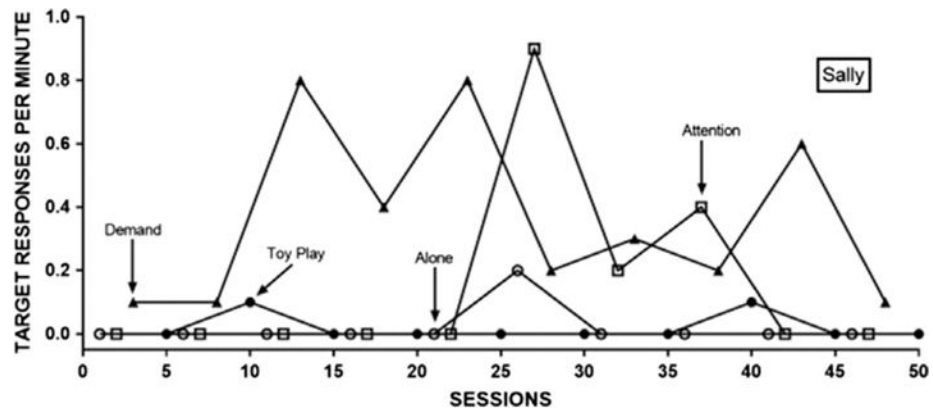
Figure 1.
Example of a functional analysis in which structured visual-inspection criteria (Roane et al., 2013) would have identified the function of destructive behavior sooner than the total number of sessions actually conducted had it been applied on an ongoing basis (recreated from Hagopian, Toole, Long, Bowman, and Lieving, 2004, with permission from the corresponding author and the journal).
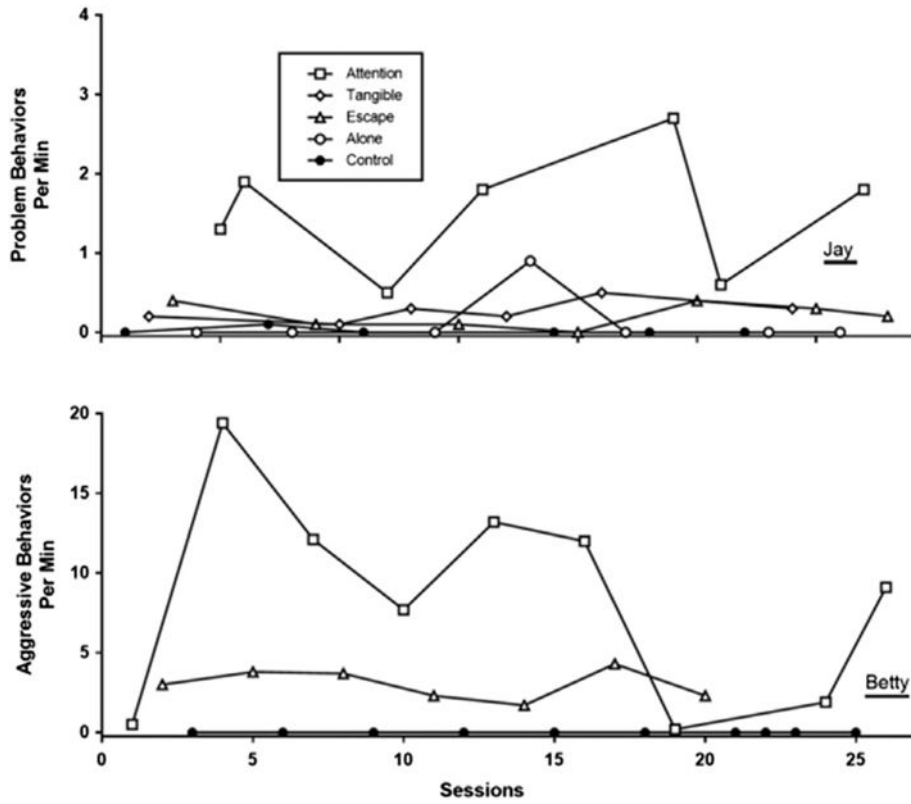
Figure 2.
Examples of extended analyses with at least six sessions per condition that we evaluated in the present study (recreated from Hanley, Piazza, Fisher, & Maglieri, 2005, with permission from the corresponding author and the journal).
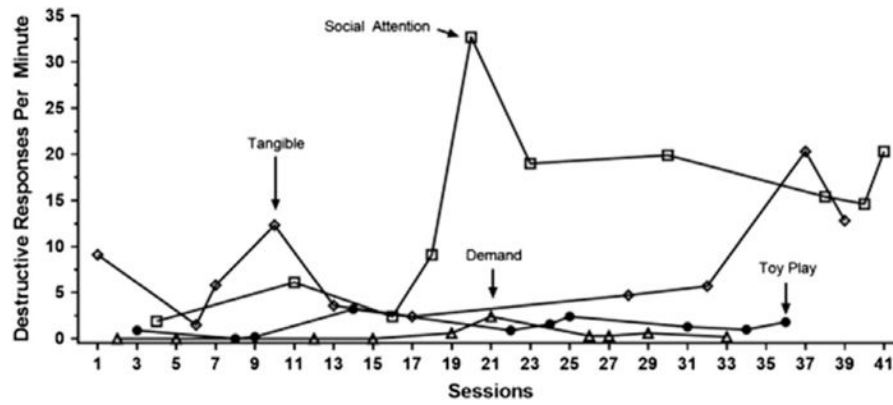
Figure 3.
Exemplary case of a disagreement between OVI and PHAI in which the authors provided an incomplete analysis of the FA results. In this case, OVI and PHVI identified attention and tangible functions for destructive behavior. By contrast, the authors discussed only the presence of an attention function and provided no information regarding the presence or absence of a tangible function (recreated from Fisher, Ninness, Piazza, and Owen-DeSchryver, 1996, with permission from the corresponding author and the journal).
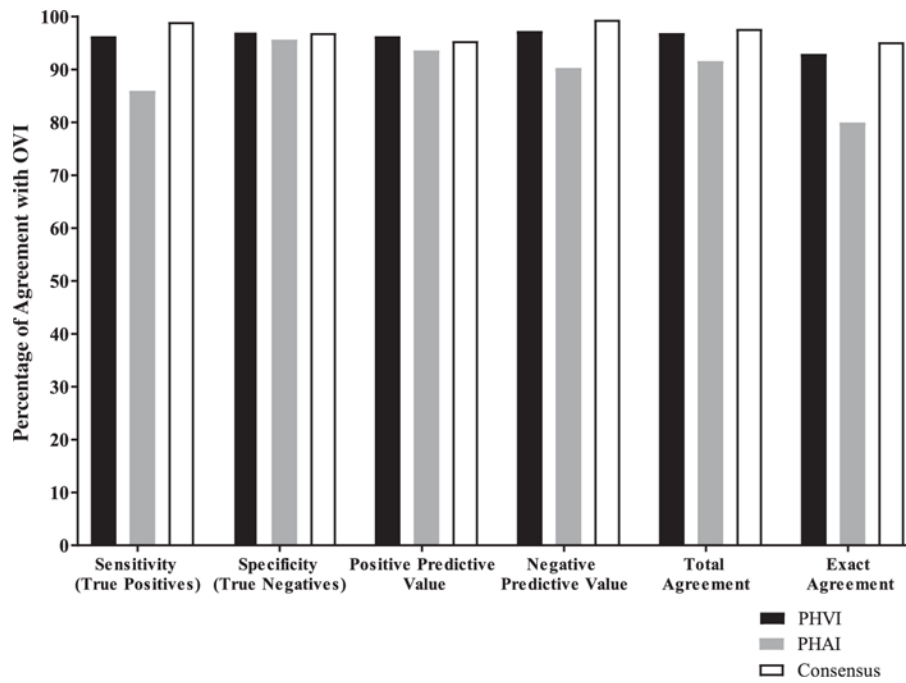
Figure 4.
Levels of agreement on measures of predictive validity regarding individual operant functions between the FA interpretations derived from OVI and the three criterion measures.
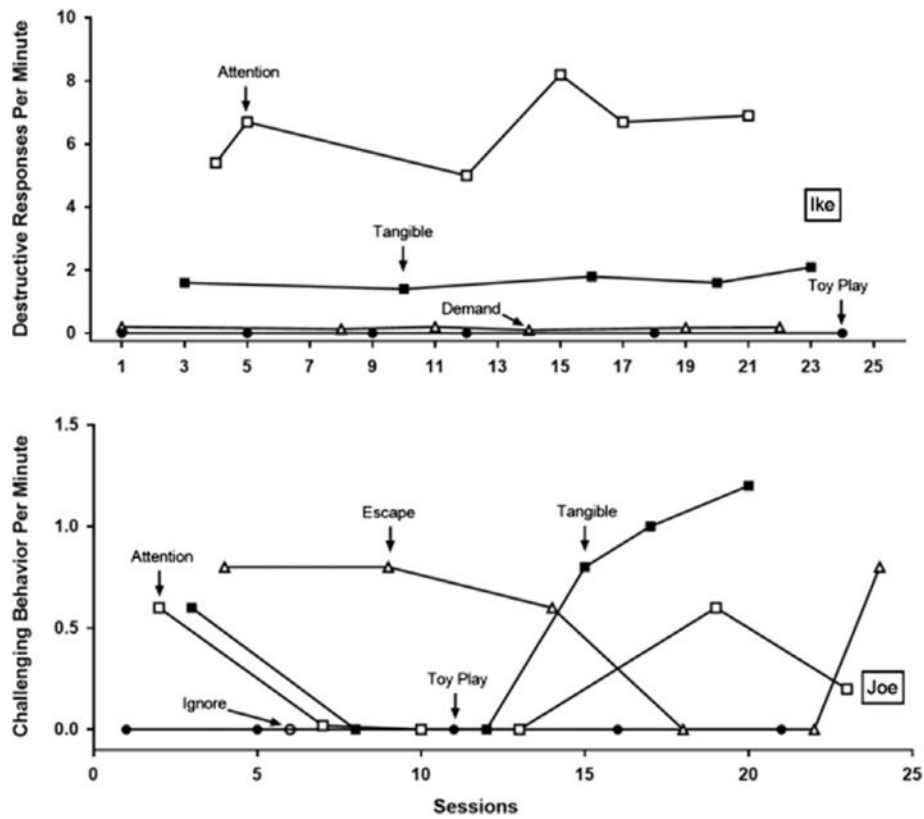
Figure 5.
An exemplar (top panel) of an exact agreement between all three interpretative methods (PHAI, PHVI, and OVI), which determined attention and tangible as the functions of destructive behavior (recreated from Piazza et al., 1999, with permission from the corresponding author and the journal). An example (bottom panel) of a disagreement between PHAI, PHVI, and OVI. PHAI identified a tangible and escape function of challenging behavior, PHVI identified a tangible function only, and OVI identified an escape function only (recreated from Falcomata, Muething, Gainey, Hoffman, & Fragale, 2013, with permission from the corresponding author and the publisher).
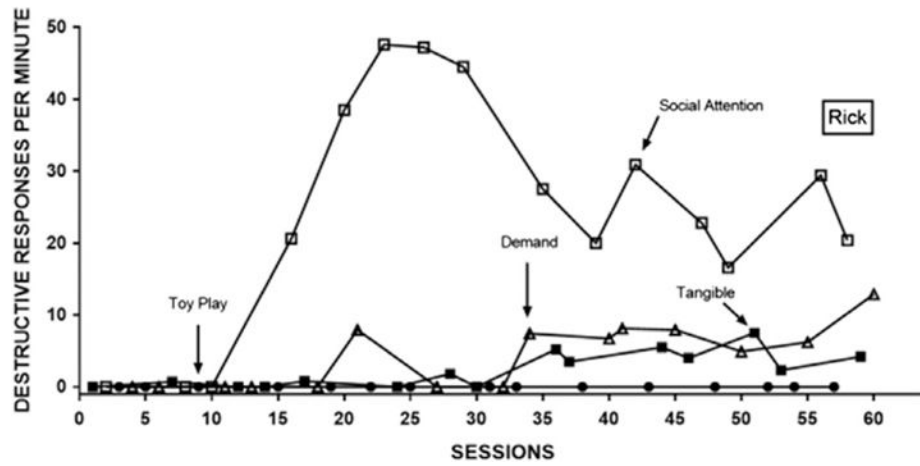
Figure 6.
Exemplary case of a disagreement between OVI and the two criterion measures. In this case, PHVI and PHAI identified attention, escape, and tangible functions for destructive behavior. In contrast, OVI identified only an attention function because levels of responding during the demand and tangible conditions increased toward the second half of the FA, after OVI terminated (recreated from Hanley, Piazza, and Fisher, 1997, with permission from the corresponding author and the journal).
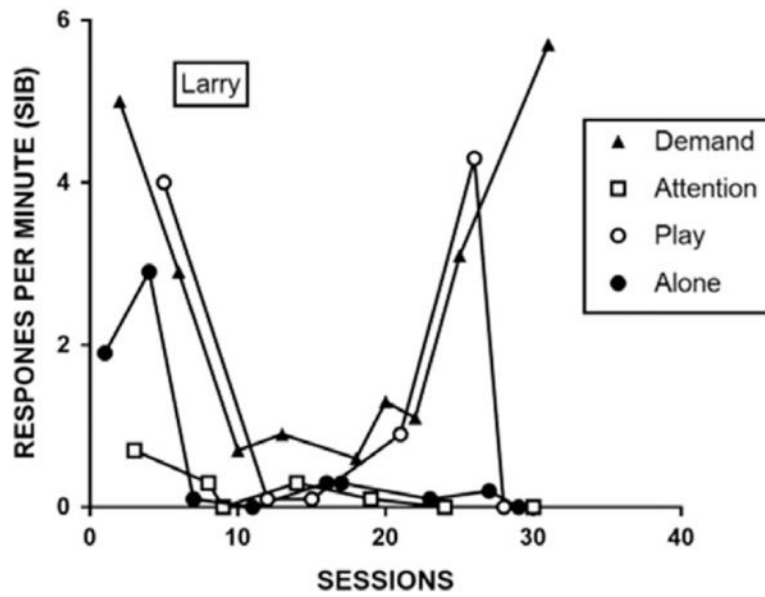
Figure 7.
Exemplary case of a disagreement wherein the authors identified an escape function of the target behavior, but OVI and PHVI concluded that the results were undifferentiated (recreated from Smith, Iwata, Goh, and Shore, 1995, with permission from the corresponding author and the journal).
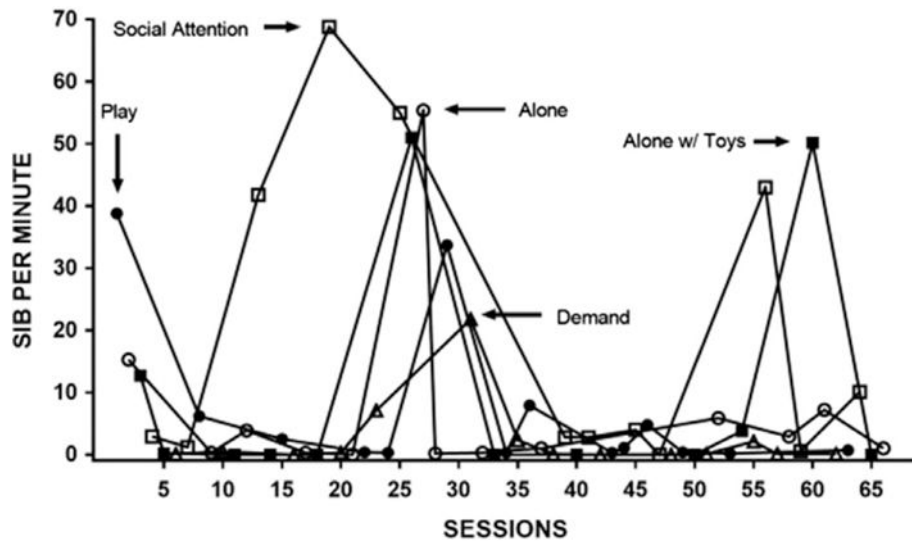
Figure 8.
The one case of an agreement between PHAI and OVI, and disagreement with PHVI. PHAI and OVI identified an attention function of self-injurious behavior, whereas PHVI concluded the FA was undifferentiated (recreated from Adelinis, Piazza, Fisher, & Hanley, 1997, with permission from the corresponding author and the journal).

Table 1

Levels of exact agreement between OVI and the three criterion measures by function.

Behavioral Function	Criterion Measure	Number and Percentage of Exact Agreement with OVI
Number of Analyses using PHVI		
Social Positive (Attention)	20	19 (95.0%)
Social Positive (Tangible)	14	13 (92.9%)
Social Negative	15	15 (100.0%)
Automatic	13	13 (100.0%)
Multiple Control	33	29 (87.9%)
Undifferentiated	5	4 (80.0%)
Total	100	93 (93.0%)
Number of Analyses using PHAI		
Social Positive (Attention)	23	19 (82.6 %)
Social Positive (Tangible)	18	13 (72.2%)
Social Negative	17	15 (88.2%)
Automatic	16	14 (87.5%)
Multiple Control	23	18 (78.3%)
Undifferentiated	3	1 (33.3%)
Total	100	80 (80.0%)
Number of Analyses using PHVI and PHAI Consensus		
Social Positive (Attention)	19	18 (94.7%)
Social Positive (Tangible)	13	13 (100.0%)
Social Negative	15	15 (100.0%)
Automatic	13	13 (100.0%)
Multiple Control	21	18 (85.7%)
Undifferentiated	2	2 (100.0%)
Total	83	79 (95.2%)

Table 2

Levels of agreement on the presence and absence of functions based on the FA interpretations derived from OVI and the three criterion measures.

OVI	Criterion Measure	
	PHVI Interpretation	
	Function Present	Function Absent
Function Present	131 (a)	5 (b)
Function Absent	5 (c)	180 (d)
	PHAI Interpretation Function Present	
	Function Absent	Function Absent
Function Present	117 (a)	19 (b)
Function Absent	8 (c)	177 (d)
	PHVI and PHAI Consensus Interpretation	
	Function Present	Function Absent
Function Present	103 (a)	1 (b)
Function Absent	5 (c)	156 (d)

Table 3

Number of FAs for which all interpretative methods, two interpretative methods, and no interpretative methods produced an exact agreement.

Interpretative Method	Number of Applications
Exact Agreement Across PHVI, PHAI, and OVI	79
Exact Agreement Between PHVI and PHAI, but not OVI	4
Exact Agreement Between PHVI and OVI, but not PHAI	14
Exact Agreement Between PHAI and OVI, but not PHVI	1
No Exact Agreement Across PHVI, PHAI, or OVI	2