Reliability, Validity, and Accuracy of the Intervention Selection Profile–Function: A Brief Functional Assessment Tool

Stephen P. Kilgus[1], Nathaniel P. von der Embse[2], Katie Eklund[1], Jared Izumi[3], Casie Peet[2], Lauren Meyer[1], & Crystal N. Taylor[3]

University of Wisconsin-Madison[1]

University of South Florida[2]

University of Missouri[3]

School Psychology Quarterly

©2019 American Psychological Association (APA). This manuscript is not the copy of record and may not exactly replicate the final version. The version of record is available at http://dx.doi.org/10.1037/spq0000325

Preparation of this article was supported by funding provided by the Institute for Education Sciences, U.S. Department of Education (R305A170284) to the first, second, and third authors. Opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education and such endorsements should not be inferred. Correspondence regarding this article should be addressed to Stephen Kilgus at the University of Wisconsin-Madison, Department of Educational Psychology, Madison, WI 53706; email: skilgus@wisc.edu

Abstract

The purpose this study was to evaluate the reliability, validity, and accuracy of scores from the *Intervention Selection Profile–Function (ISP-Function)*: a brief functional assessment tool founded upon Direct Behavior Rating (DBR) methodology. Participants included 34 teacher-student dyads. Using the ISP-Function, teachers rated the extent to which students exhibited disruptive behavior, as well as the frequency with which disruptions were met with four consequences. Ratings were completed across three 10-minute sessions, during which a research assistant also collected systematic direct observation (SDO) data regarding the same behavior and consequences. Results indicated adequate temporal reliability (≥.70) was attained for the Adult Attention and Peer Attention target across the three ratings; in contrast, up to 8-18 data points would be needed to achieve adequate reliability across the remaining targets. Findings further suggested that while ISP-Function ratings of Disruptive Behavior, Adult Attention, and Peer Attention were moderately to highly correlated with SDO data, correlations were in the low range for the Access to Items/Activities and Escape/Avoidance targets. Finally, analysis of difference scores showed that on average, mean ISP-Function scores fell within only 0.33 to 1.81 points of mean SDO scores (on the 0-10 DBR scale). Agreement coefficients indicative of exact score agreement were less consistent, suggesting accuracy ranged from poor to substantial. Results are promising, but future research is necessary to support applied ISP-Function use.

Impact Statement: Study results support the continued validation of the *Intervention Selection Profile – Function*, with findings speaking to the reliability, validity, and accuracy of three of five scales within the tool. The availability of such tools enhances the likelihood of schools collecting functional assessment data to inform the implementation of Tier 2 interventions.

Reliability, Validity, and Accuracy of the Intervention Selection Profile–Function: A Brief Functional Assessment Tool

Comprehensive functional behavior assessment (FBA) is a multi-method and multi-informant process for gathering information that can be used to address problem behavior (O'Neill, Albin, Store, Horner, & Sprague, 2015). Meta-analyses have found that function-based interventions informed by FBA findings are associated with medium-sized effects (Bruni et al., 2017; Gage, Lewis, & Stichter, 2012). These results have been found irrespective of disability status or location of intervention (i.e., self-contained special or general education classroom; Bruni et al., 2017). Relatedly, research has found that when interventions are delivered in a uniform fashion across a wide range of students, the function of student problem behavior can moderate intervention effectiveness (Gage et al., 2012; Walker, Chung, & Bonnet, 2018). For instance, multiple studies have indicated Check In/Check Out, a common Tier 2 intervention, is more effective for students whose behavior functions to gain attention, as compared to students whose behavior functions to escape aversive stimuli (e.g., McIntosh, Campbell, Carter, & Dickey, 2009). Fortunately, researchers have also shown that by applying function-based adaptations of these interventions in accordance with FBA findings, larger effects can be attained (Klingbeil, Dart, & Schramm, 2018).

FBA at Tier 2

Within schools, FBA and function-based intervention have commonly been seen as only relevant to students exhibiting more severe problem behavior at Tier 3 and special education (Young, Andrews, Hayes, & Valdez, 2018). More recently, scholars have called for FBAs to be conducted in relation to less intensive problem behaviors, including those addressed through Tier 2 supports (Reinke, Herman, & Stormont, 2013). Some have raised concerns regarding the feasibility of comprehensive FBAs at Tier 2, as it is expected that 10-15% of students within a school might require support at this level (Cheney, Flower, & Templeton, 2008). In response, scholars have further suggested that comprehensive FBAs might not be necessary at Tier 2; rather, it might be appropriate to base function-related decisions on data from a single FBA tool (McIntosh, Brown & Borgmeier, 2008). Justification for the use of a single tool is founded in recognition of the rather low-stakes nature of Tier 2 decisions and the presumed simplicity of problem behaviors at this level (relative to what is typical of more intensive cases at Tier 3 or within special education; Dunlap & Kern, 2018).

Should a school choose to employ a single FBA tool at Tier 2, a question would still remain regarding what type of tool is most appropriate. Unfortunately, a review of existing options reveals one or more limitations to many FBA tools, which would make their use at Tier 2 somewhat challenging. For example, many tools are time- and resource-intensive. To illustrate, systematic direct observation (SDO) is one of the more common FBA tools (Lloyd, Weaver, Staubitz, 2017) and typically requires a specifically trained third-party observer to be present within relevant structured and unstructured settings (e.g., classroom and playground) to collect data across multiple sessions. Though this observational process yields highly objective and low inference data (as they are collected during the time and setting when the behavior in question is displayed), it can nevertheless be challenging to coordinate and implement on a large scale, such as at Tier 2. Thankfully, prominent scholars have suggested that Tier 2 cases might not require the collection of SDO data, and could instead be based upon indirect data collection tools, such as semi-structured interviews and rating scales (Dunlap & Kern, 2018).

Yet, there are limitations to many indirect FBA tools. For example, multiple studies have revealed available rating scales are characterized by inconsistent or inadequate performance. To illustrate, a recent investigation found the *Functional Analysis Screening Tool* (Iwata & DeLeon, 1995), a function-specific rating scale, identified the correct function of behavior in only 63.8% of 69 cases (Iwata, DeLeon, & Roscoe, 2013). Additional studies have found the *Motivation Assessment Scale* (Durand & Crimmins, 1988), another rating scale, to yield inconsistent inter-rater reliability, resulting in variance in functional decisions across raters (Zarcone, Rodgers, Iwata, Rourke, & Dorsey, 1991). Beyond these specific tools, more general concerns might be raised in relation to indirect data sources. When collecting indirect data, informants (e.g., teachers, parents) are required to describe behavior the student has exhibited over the past days, weeks, and months. Such extended latency between the actual behavior and informant recall indicates indirect data are likely to be rather subjective, high inference, and open to a number of common rater biases (e.g., halo, contrast; Christ, Riley-Tillman, & Chafouleas, 2009).

In summary, many existing FBA tools are associated with limitations that impede their utility at Tier 2, including time- and resource-intensiveness or subjectivity. As such, there is a need to develop and validate novel tools that shed these limitations and thus enhance the potential for evidence-based FBA at Tier 2. Researchers have recently begun to consider whether Direct Behavior Rating (DBR) methodology might serve as an appropriate basis of such tools.

Direct Behavior Rating

DBR is a broader assessment methodology, representing a hybrid of both SDO and behavior rating scale methods (Chafouleas, 2011). Like SDO, DBR requires an individual to observe a student's behavior within a pre-specified time and setting (e.g., small-group reading instruction, 2:15-3:00pm). Like a behavior rating scale, the individual indicates the extent to which a behavior was observed during that time through a brief rating. Though the scale through which ratings are completed will vary across specific DBR tools, the most common is the 11-

point unipolar graphic rating scale, which increases in increments of 1 and is anchored at the beginning (0; 0% Never), middle (5; 50% Sometimes), and end (10; 100% Always). Through its use of data collectors commonly present within the data collection setting (e.g., classroom teachers), DBR tends to be more feasible than SDO. Furthermore, by requiring the collection of data immediately following the display of behavior, DBR tends to be lower inference than typical behavior rating scale scores, which correspond to behaviors displayed during the previous weeks or even months.

To date, the majority of DBR research has examined the measure's capacity to evaluate behaviors alone (e.g., Riley-Tillman, Chafouleas, Sassu, Chanese, & Glazer, 2008). Studies have suggested DBR scores are valid and accurate indicators of student behavior (Kilgus, Riley-Tillman, Stichter, Schoemann, & Bellesheim, 2015). For example, Riley-Tillman et al. (2008) found that DBR estimates of student on-task and disruptive behavior were highly and statistically significantly correlated with SDO estimates of these same behaviors, when both sets of data were collected within the same setting and time period (mean r = .81-.87). Study results also demonstrated that the majority of mean DBR estimates of behavior fell within 1-2 points of mean SDO estimates along the 0-10 DBR scale. This latter finding supported DBR accuracy, suggesting that broader conclusions regarding levels of on-task and disruptive behavior were consistent across the two methods.

Additional research has supported the reliability of DBR scores. For instance, Kilgus et al. (2015) examined the temporal reliability of DBR data, defined as the stability of score estimates over time (Cicchetti, 1994). Analyses examined how many data points would have been needed to achieve a minimum reliability of .80, a common threshold for adequate score reliability (Chafouleas et al., 2010). Findings suggested that across DBR targets, only 2-6 data points were needed to reach this threshold, thus speaking to the feasibility of DBR data collection within applied settings.

DBR for Functional Assessment

Given its promising performance in evaluating behavior, researchers have begun to consider whether DBR can generate defensible evidence of other types of targets. For example, multiple studies have examined whether DBR could be used to examine the consequences that follow behavior, with an interest in evaluating whether DBR is capable of a broader role within FBA. An initial study included 213 undergraduate student participants who viewed and rated video clips of students within a classroom setting (Kilgus, Kazmerski, von der Embse, & Taylor, 2017). Immediately following the end of each clip, participants used DBR scales to rate (a) the percentage of time a target student exhibited disruptive behavior; and (b) the percentage of disruptive instances that were met with each of four consequences, including adult attention, peer attention, escape/avoidance, and access to items/activities. This latter rating type was intended to be interpreted in a manner consistent with conditional probabilities, which are commonly evaluated in determining the likely function(s) of a problem behavior (Martens, DiGennaro, Reed, Szczech, & Rosenthal, 2008). Results indicated that following training with opportunities to practice DBR ratings with performance feedback, participants generated accurate ratings of disruptive behavior and its various consequences. More specifically, participants were able to generate ratings within 0 to 2 points of video clip "true scores," which were attained through SDO. A review of SDO data relative to mean DBR scores across participants indicated one would likely make the same conclusion regarding the level of disruptive behavior and its function across the two assessment methods.

In a follow-up study, South (2017) evaluated the validity of DBR function-based ratings within an in-vivo format. More specifically, the researcher compared (a) paraprofessional-completed DBR ratings of disruptive behavior and its consequences to (b) external observer-collected SDO scores. A large and statistically significant correlation was noted between DBR and SDO estimates of disruptive behavior ($r = .89$). Correlations between the two methods were comparatively smaller for the consequence targets, with coefficients ranging from -.05 to .47. Notably, of the consequence correlations, the highest coefficient was documented for Adult Attention, which was also statistically significant. The remaining correlations were <|.10|. Though many reasons for these lower correlations might exist, one particularly probable reason pertains to floor effects associated with the Escape/Avoidance and Access to Items/Activities targets. Notably, across the broader study, Escape/Avoidance was not documented once through SDO, whereas Access to Items/Activities was documented only twice. Such restricted variance will invariably attenuate correlations and make it difficult to evaluate the performance of these targets within this context.

Purpose of the Study

In summary, initial research has yielded preliminary but inconsistent support for the use of DBR in collecting functional assessment information. The purpose of this study was to address certain limitations to this existing research, particularly that related to South (2017). First, the South (2017) study included only seven participants, raising concerns regarding sampling error, limited external validity, and restricted statistical power. Second, the study was conducted in a private residential treatment facility with paraprofessional raters. Though research conducted in such a setting is certainly important, findings might have limited applicability to more typical public-school settings. Furthermore, previous research has suggested the defensibility of classroom teacher-collected DBR data might exceed that of paraprofessional-collected data (Chafouleas et al., 2010). Thus, it would appear additional research is needed to understand how DBR functional assessment data might perform when collected within typical school settings by classroom teachers. Third, the author did not evaluate the temporal reliability of formatively-collected DBR data points. This is unfortunate, as evidence of score reliability has implications for data collection guidelines, informing the number of data points that should be collected (Kilgus et al., 2015).

The purpose of this study was to build upon the nascent evidence for DBR in functional assessment while addressing the aforementioned limitations to the South (2017) investigation. The particular tool under investigation was the newly titled *Intervention Selection Profile–Function* (ISP-Function), a specific DBR-based tool incorporating the disruptive behavior and consequence targets described above. Three research questions were posed. First, to what extent do DBR targets exhibit temporal reliability? Based upon prior DBR research (e.g., Chafouleas et al., 2010), it is anticipated that the three DBR data points collected for each student within each target will not exhibit adequate reliability for low-stakes decisions (≥.70; Salvia, Ysseldyke, & Bolt, 2007), like the determination of behavioral function at Tier 2. Thus, it is further hypothesized findings will suggest that 5-15 data points will be necessary to achieve adequate reliability (Chafouleas et al., 2010).

Second, to what extent do ISP-Function ratings exhibit criterion-related concurrent validity, as compared to SDO estimates? Based upon previous DBR research (e.g., Riley-Tillman et al., 2008), it was hypothesized ISP-Function ratings would be moderately to highly positively correlated ($\rho > .30$ and .50, respectively; Cohen, 1988) with SDO-based (a) estimates of disruptive behavior, as indicated by percentage of interval data; and (b) conditional probability statistics, indicative of the functional relation between disruptive behavior and various consequences. Third, to what extent are ISP-Function scores accurate estimates of SDO estimates? Based upon previous research (Kilgus et al., 2017), it was hypothesized ISP-Function disruptive behavior and consequence ratings would be similar in terms of level (i.e., *M*) relative to SDO estimates of disruptive behavior (i.e., percentage of intervals) and consequences (i.e., conditional probability statistics), as indicated by absolute difference scores. It was further expected that ISP-Function scores would demonstrate fair to moderate agreement with SDO scores, as indicated change-corrected agreement coefficients.

Method

Participants

Teachers were recruited from eight elementary (Kindergarten through fifth grade) schools across two sites, including one in the Midwest and another in the Southeast. Upon receiving university and school district Institutional Review Board approvals, a recruitment letter was distributed to the elementary school staff describing the study. Interested teacher participants were then asked to identify a student in their classroom that commonly exhibited disruptive behavior. Teachers were instructed to not identify students engaging in highly problematic or dangerous disruptions (e.g., aggressive acts), given the study's focus on students at the Tier 2 level. Teachers then sent home a consent form to the identified student's parents or guardians. Once parental consent was received, the teacher-student dyad was enrolled in the study. A total of 34 teacher-student dyads were enrolled in this study across the two sites (Southeast *n* = 15; Midwest *n* = 19). A majority of teacher participants were female (33 out of 34) and White or Caucasian (82%). Additional participant demographic information is provided in Table 1.

Measures

Intervention Selection Profile-Function. The ISP-Function is a brief five-target measure founded upon DBR methodology. It is intended to measure the degree to which a student engages in an operationally defined problem behavior and the frequency in which the target behavior is met with one of four consequences, including adult attention, peer attention, escape/avoidance, and access to tangibles or activities. *Disruptive behavior* was defined as a student action that interrupts regular school or classroom activity (e.g., out of seat, fidgeting, playing with objects, acting aggressively, and talking/yelling about things that are unrelated to instruction). *Adult attention* was defined as positive, negative, or neutral adult reaction that can be verbal or nonverbal (e.g., reprimand, redirection to work). *Peer attention* was defined as positive, negative, or neutral peer reaction that can be verbal or nonverbal (e.g., talking, laughing, arguing). *Escape/avoidance* was defined as the removal or avoidance of task, activity, or performance expectations (e.g., removal of academic materials, permission to leave room). Lastly, *access to tangibles/activities* was defined as acquisition of items or activities (e.g., toys, food, prizes, preferred tasks).

The ISP-Function is completed by a teacher or individual who (1) is familiar with the student, (2) has the opportunity to observe the student in a classroom environment, and (3) is able to complete the measure soon after the specified observational period (e.g., 10:00 to 10:30 during math class). Prior to collecting ISP-Function data, the teacher (in consultation with special educators or other related support personnel) must select the time and setting within which they will observe each target student. Immediately following these rating periods, the teacher completes five ratings. The first rating corresponds to the user's perception of the percentage of total time the student exhibited problem behavior during the rating period. Ratings 2-5 correspond to the percentage of problem behaviors that were followed by each consequence.

ISP-Function data are scored and interpreted using a multi-step process. First, the mean of scores is calculated within each ISP-Function target, yielding a more generalizable estimate of each behavior and consequence. Resulting problem behavior scores represent estimates of the percentage of time students engaged in the problem behavior, whereas resulting consequence scores represent the percentage of problem behaviors met with each consequence. Second, means are plotted on a bar graph, with ISP-Function items on the *x*-axis and percentage scores on the *y*-axis. The higher the mean level of problem behavior, the more likely that a function-based intervention would be appropriate. Third, to determine the function of a student's behavior and thus support intervention matching, consequence mean scores are interpreted in accordance with conventions for conditional probability analysis, with higher values representing more likely behavior functions (Eckert, Martens, & DiGennaro, 2005).

Systematic Direct Observation. SDO served as the criterion data source within this investigation. Observations were founded upon a behavior-consequence (BC) partial interval recording procedure. Each observation lasted 10 minutes and was divided into 20-sec intervals. SDO also considered the same behavior and consequence targets (with corresponding operational definitions) as rated by the teacher using the ISP-Function. In accordance with the partial interval procedure, an interval was marked as including disruptive behavior if the student was disruptive at any point during the interval. The observer also noted which consequence(s), if any, occurred following the disruptive instance within each disruptive interval. Two types of scores were derived following each SDO. First, behavior SDO scores represented the percentage of total intervals within which disruptive behavior was recorded. Second, four consequence SDO scores were calculated, including one per consequence target. Each consequence score represented a conditional probability, defined as the percentage of disruptive intervals that included the consequence in question (Eckert et al., 2005). A second observer was present for 14% of all observations for the purpose of evaluating inter-observer agreement (IOA). IOA within each observation represented the percentage of observed intervals coded identically. The mean IOA across observations was 87% for disruptive behavior (range = 70-97%) and 97% for consequence targets (range = 80-100%).

SDO is considered the gold standard of behavior assessment methods, with a rich psychometric base and history of use within the literature (Chafouleas, Kilgus, Riley-Tillman, Jaffery, & Harrison, 2012). BC recording in particular has also received support across several investigations, with findings supporting the accuracy of BC recording data in identifying behavioral function (Lewis, Mitchell, Harvey, Green, & McKenzie, 2015).

Procedures

Multiple research assistants helped to drive study procedures. Each of these assistants was a doctoral student in school psychology who had at least one year of practicum experience and had completed coursework in applied behavior analysis and behavior assessment and intervention. Prior to data collection activities, research assistants were trained in data collection procedures by the project investigators. First, assistants were provided an overview of the ISP-Function, with instructions regarding how assistants should train teachers to use the tool. Second, research assistants were trained to complete the SDO procedure. Following an overview of the operational definitions and partial interval recording procedure, research assistants were provided the opportunity to practice using the SDO form while viewing a series of five-minute videos of students within a classroom setting. SDO practice continued until all research observers' inter-observer agreement with the principal investigator conducting the training met or exceeded a minimally acceptable level (i.e., Cohen’s kappa = .60; Kratochwill et al., 2010).

Next, each research assistant was assigned multiple teacher-student dyads with which they would be working. Research assistants then interacted with each teacher across a series of meetings. During *Meeting 1*, the research assistant confirmed disruptive behavior was indeed a concern for the student, while also determining the times and settings when disruptive behavior was most common (e.g., Math, independent seatwork, 11:00-11:45). During *Meeting 2*, the research assistant trained the teacher to complete the ISP-Function using a detailed training script. Training procedures included an introduction on the basics of behaviorism and FBA, a review of the ISP-Function form, and practice rating video clips of student behavior. “True scores” for each video had been determined through SDO in advance, permitting the provision of feedback to each teacher regarding their rating accuracy. The average length of ISP-Function trainings was 36 minutes.

Following *Meeting 2*, the teacher collected ISP-Function data across three 10-minute classroom observation sessions. The research assistant was present during each session to collect SDO data throughout the same three 10-min sessions, permitting direct comparability of ISP-Function and SDO data. At the beginning of each session, the research assistant would note to the teacher that they should begin to observe the student’s behavior and its associated consequences. The research assistant then began their SDO procedure. A timer application on each research assistant’s phone supported observations, vibrating at the end of each 20-sec interval (to indicate the assistant should record) and at the end of 10-min period. At the end of the 10-min session, the research assistant ended their observation and approached the teacher to complete their rating using a paper-based ISP-Function form. The teacher quickly recorded their ratings and returned to their typical teaching activities.

Data Analysis Plan

Research question 1. In a manner consistent with Kilgus et al. (2015), we examined the temporal reliability of ISP-Function data in accordance with recommendations from Shrout and Fleiss (1979). First, intraclass correlation (ICC) coefficients were calculated via a two-level unconditional hierarchical linear model, wherein repeated ISP-Function data points (Level 1) were nested within student participants (Level 2). ICCs were computed as the ratio of between-student variance to total variance (between-student variance + within-student variance). Next, ICCs were used to generate reliability estimates using the following formula:

$$r_{xx} = \frac{k * ICC}{(1 + (k - 1) * ICC}$$

Where *k* is the number of observations and ICC is equal to the ICC described above. For each analysis, *k* was set to 3, as this represented the number of ISP-Function ratings collected for each student participant. Next, we computed the minimum number of data points ($\min_k$) required to achieve reliabilities of .70 by re-expressing the equation above to solve for *k* and fixing $r_{XX} = .70$.

Research question 2. No missing data were noted across all ISP-Function and SDO targets; thus, it was not necessary to employ a missing data handling technique. Next, we derived mean scores across sessions within each ISP-Function and SDO target. We then conducted a linear transformation of SDO values by dividing each mean SDO score (0-100) by 10 so that it was on the same scale as ISP-Function scores (0-10). This transformation was done to enhance the comparability of scores from the two methods, as required for Research Question 3 analyses.

Next, to address Research Question 2, we calculated Spearman's rho (ρ) non-parametric correlation coefficients comparing ISP-Function and SDO in terms of individual session scores and across-session mean scores. The decision to employ non-parametric correlations was made in consideration of descriptive data (see Table 2), which indicated SDO data exhibited high levels of both skew and kurtosis and thus did not meet the distributional assumptions of parametric analyses. Correlation values were compared to common interpretive thresholds of small (>.10), medium (>.30), and large (>.50; Cohen, 1992).

Research question 3. The accuracy of ISP-Function scores was evaluated via two statistical procedures. First, we computed difference scores comparing ISP-Function and SDO mean scores within each student participant. Difference scores (x_{DIFF}) were calculated by computing the absolute value of the difference between each student's mean ISP-Function and SDO scores ($x_{\mathrm{DIFF}}= |x_{\mathrm{ISP}} - x_{\mathrm{SDO}}|$). Per this scoring approach, lower scores were interpreted as indicative of greater accuracy. The decision to examine absolute values (as opposed to non-absolute values) followed our interest in not allowing otherwise positive and negative difference scores to cancel one another out across students, as this would reduce mean difference score estimates and overestimate accuracy. Five difference scores were computed for each student, comparing mean ISP-Function to SDO scores within each target of interest. The mean difference score was then examined within each target. Furthermore, we evaluated the cumulative percentage of students falling within various difference score levels. Specifically, we examined the percentage of students whose mean ISP-Function scores fell within 0, 1, and 2 points of their mean SDO scores on a 0-10 DBR scale. This approach is consistent with prior work evaluating the accuracy of DBR data, wherein it was of interest to examine the extent to which DBR ratings approximated SDO scores, even if not yielding the same exact value (Riley-Tillman et al., 2008).

Second, Gwet's agreement coefficient (AC), as well as corresponding 95% confidence intervals (CI-95) and *p* values, was computed to examine agreement between ISP-Function and

SDO scores after accounting for chance. AC represents an alternative to the more commonly applied kappa statistic, which is overly sensitive to behavioral base rates and the magnitude of raters' classification probabilities (Viera & Garrett, 2005). Given the potential for disruptive behavior and various consequences to appear infrequently within this sample, the AC statistic was considered to afford a more accurate depiction of ISP-Function accuracy. Prior to analyses, mean scores were rounded to the nearest whole number. Accuracy was then defined as the extent to which both methods yielded the same rounded score. Such an approach is consistent with prior DBR accuracy work, which defined accuracy as perfect agreement between DBR and SDO in terms of whole number on the 0-10 DBR scale (LeBel, Kilgus, Briesch, & Chafouleas, 2010). AC values range between 0 and 1, with values closer to one representing stronger agreement. AC values less than .20 were classified as poor, .21–.40 fair, .41–.60 moderate, .61–.80 substantial, and >.80 nearly perfect (Klein, 2018; Landis & Koch, 1977). The null hypothesis of method independence was rejected if the *p* value associated with each AC statistic was less than .05.

The two aforementioned approaches to evaluating accuracy are associated with varying levels of stringency. The first approach is less stringent, as it involves examining the extent to which ISP-Function and SDO scores tend to fall around each other, as evidenced by bands of difference scores (e.g., ≤1 and ≤2). The second is more stringent, permitting the examination of exact agreement between the two methods in terms of whole score of the DBR 0-10 scale. When taken together, we believe the two approaches afford more holistic depiction of the extent to which ISP-Function and SDO agreed regarding (a) the level of student disruptive behavior and (b) the frequency with which behaviors were met with varying consequences. To note, all analyses were conducted using Stata Version 15.1. In addition, AC statistics were calculated using the *kappaetc* command (Klein, 2018).

Results

Research Question 1

See Table 3 for an overview of reliability findings. Results suggested the three collected data points were sufficient to achieve adequate low-stakes reliability (≥.70) for the Adult Attention and Peer Attention targets. The remaining three targets did not demonstrate adequate reliability, with all values falling below this threshold (range = .29-.46). Follow-up analyses suggested that to achieve reliabilities of .70, it would be necessary to collect 8.18 data points for Disruptive Behavior, 14.95 for Escape/Avoidance, and 17.50 for Access to Items/Activities.

Research Question 2

See Table 4 for an overview of correlations between ISP-Function and SDO scores. Mean ISP-Function disruptive behavior scores were highly correlated ($\rho > .50$) with SDO scores. Relations between mean ISP-Function scores and SDO-based conditional probability statistics varied across targets. Correlations were in the high range for both Adult Attention and Peer Attention targets. In contrast, correlations fell in the low ($\rho > .10$) range for Escape/Avoidance and Access to Items/Activities. Interestingly, correlations were in the opposite direction of expectations for the Access to Items/Activities target. A review of descriptive statistics (see below) suggested these latter two sets of low correlations might have been influenced by floor effects and a restriction of range in ISP-Function and SDO data.

See Table 4 for a summary of correlations between ISP-Function and SDO scores across individual data collection sessions. Correlations were consistently in the moderate or high range across Disruptive Behavior, Adult Attention, and Peer Attention targets. As with mean score correlations, individual session correlations were in the low range across the Escape/Avoidance and Access to Items/Activities targets. Interestingly, it was not possible to calculate correlations for the Access to Items/Activities target across sessions two and three; this was a result of all SDO values being equal to zero across participants for these sessions.

Research Question 3

See Figure 1 for a summary of mean absolute difference scores across participants and sessions. On average, ISP-Function and SDO Disruptive Behavior scores differed by only 1.01 points, or roughly one point on the 0-10 DBR scale. Mean consequence scores ranged between 0.33 (Access to Items/Activities) and 1.81 (Peer Attention); in other words, on average, ISP-Function and SDO scores tended to fall within 1-2 points of each other on the 0-10 DBR scale. Further review of descriptive statistics indicated mean difference scores were ≤1 for 61.8% of students for Disruptive Behavior, 47.1% for Adult Attention, 44.1% for Peer Attention, 67.6% for Escape/Avoidance, and 85.3% for Access to Items/Activities. Furthermore, mean difference scores were ≤ 2 for 82.4% of students for Disruptive Behavior, 64.7% for Adult Attention, 64.7% for Peer Attention, 76.5% for Escape/Avoidance, and 100.0% for Access to Items/Activities.

Next, AC statistics were computed in examining the extent to which ISP-Function and SDO methods agreed in terms of rounded mean score across participants. AC fell in the fair range and was statistically significant for the Disruptive Behavior target (AC = .30 [.11–.50], *p* = .003). The AC values fell in the poor range and were non-statistically significant for Adult Attention (AC = .11 [-.05–.27], *p* = .173) and Peer Attention (AC = .13 [-.03–.29], *p* = .117). In contrast, AC values were statistically significant and fell in the moderate and substantial ranges, respectively, for Escape/Avoidance (AC = .50 [.29–.71], *p* < .001) and Access to Items/Activities (AC = .74 [.55–.92], *p* < .001).

Discussion

Research has substantiated the value of functional assessment data in the delivery of effective, individualized, and appropriate behavioral interventions for students with a variety of behavioral needs (Filter & Horner, 2009; Ingram, Lewis-Palmer, & Sugai, 2005). As such, identifying tools that allow for an examination of behavioral function, while also demonstrating qualities of psychometric defensibility, is an important consideration. However, practitioners often desire assessment procedures that are efficient, especially when considering the broader needs of students when providing Tier 2 interventions. The ISP-Function demonstrates promise in that it possesses preliminary (albeit inconsistent) evidence indicative of its validity and accuracy relative to SDO, while also being quick and efficient for teachers to use when assessing students with behavioral concerns. The broader purpose of this study was to build upon this prior evidence in evaluating the suitability of ISP-Function use within general education settings.

Research Question 1

The first research question under consideration pertained to the temporal reliability of ISP-Function scores. In accordance with previous research (e.g., Chafouleas et al., 2010), it was hypothesized that the three formatively-collected scores would not exhibit sufficient reliability within any of the ISP-Function targets. Hypotheses were supported for the Disruptive Behavior, Escape/Avoidance, and Access to Items/Activities targets (reliability range = 0.29-0.46). Follow-up analyses suggested across these targets, it would be necessary to collect anywhere from 8-18 data points to achieve sufficient reliability of low-stakes decisions (≥.70). Interestingly, in contrast to these three targets, both the Adult Attention and Peer Attention targets exhibited adequate reliability across the three collected data points (range = 0.72-0.75). Furthermore, follow-up analyses indicated it would be necessary to collect only 4-5 data points to achieve sufficient reliability for high-stakes decisions.

Collectively, the current results are in accordance with prior research, which has suggested that it would be necessary to collect 5-15 DBR data points to achieve sufficient reliability (Chafouleas et al., 2010). Such a finding is considered encouraging in the practical sense, as it rather feasible to collect this amount of data within a school setting. As noted earlier, it takes only a few seconds to complete DBR ratings like those in the ISP-Function. Furthermore, given this brevity, as well as the fact that DBR ratings are tied to specific times and settings, it is possible to collect multiple data points across consecutive days. This means it could take only a few or even one day to collect the data necessary to achieve reliable data for low-stakes functional assessment decisions at Tier 2. Despite this potential feasibility, future research should examine how to reduce the number of ISP-Function data points required to achieve adequate reliability, as the necessity of a larger number of data points for each student could prove challenging. Methods by which to enhance reliability might include the use of enhanced teacher training or the use of alternative ISP-Function target definitions.

Finally, it should be noted that the current reliability findings are likely reflective of the observed between-student variance in ISP-Function scores. The temporal reliability statistic used within this study is upwardly biased when evaluating measures with high variability across participants. As noted above, certain ISP-Function targets were consistently low for all students within this sample. Reliability will then be expectedly low for these targets, suggesting the need for additional data points to achieve sufficient reliability. Moving forward, consideration of a more diverse sample in terms of disruptive behavior and behavioral function will likely support a more accurate estimate of ISP-Function reliability under typical circumstances.

Research Question 2

As with reliability, support for ISP-Function validity was mixed across targets. Results specific to the Disruptive Behavior, Adult Attention, and Peer Attention targets were in accordance with hypotheses, as well as previous research comparing DBR ratings to SDO data (Riley-Tillman et al., 2008; South, 2017). Specifically, correlations were consistently in the moderate or high range when assessing either mean scores and individual session scores. Such findings indicate the ISP-Function is capable of affording information regarding the level of a student's disruptive behavior, as well as the extent to which such behavior is met with attention from others.

Conversely, correlations were in the low range for the Access to Items/Activities and Escape/Avoidance targets. Given the lower frequencies with which these consequences were observed, results may have been influenced by floor effects. For example, in two of the three observational sessions, SDO data indicated these consequences were never observed. These results were comparable to previous studies that also found floor effects associated with these latter two targets (South, 2017). This consistent finding of restricted variance across studies indicates the evaluation of these consequences may continue to be challenging across future investigations, particularly when applying correlational metrics. Regardless, given findings to date, there is less support for the use of the ISP-Function in evaluating whether a student's disruptive behavior is maintained by escape/avoidance or access to items/activities. It therefore remains necessary to conduct research with more behaviorally diverse samples inclusive of those exhibiting an array of functional behavior profiles.

Research Question 3

The final research question pertained to the accuracy of ISP-Function estimates relative to SDO scores. Descriptive analysis of absolute difference scores showed that on average, ISP-

Function and SDO scores differed by only one point (on a 0-10 DBR scale) for disruptive behavior. Findings were similar for the consequence targets, with mean difference scores ranging between 0.33-1.81. Further evaluation of difference scores indicated mean ISP-Function scores were within two points of mean SDO scores for 64.7-100.0% of students across the five targets. Taken together, descriptive findings align with previous research, whereby DBR behavior estimates fell within 1-2 points of SDO estimates (Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005; Riley-Tillman et al., 2008).

In contrast to difference score findings, AC statistics portrayed a somewhat inconsistent view of ISP-Function accuracy. Specifically, though AC values were statistically significant and fell in fair, moderate, or substantial range for Disruptive Behavior, Escape/Avoidance, and Access to Items/Activities (respectively), AC values were non-statistically significant and in the poor range for both Adult Attention and Peer Attention.

In summary, difference score findings suggested teachers who were otherwise engaged in instructional activities were able to monitor student behavior and the consequences that followed, and then generate ratings that closely approximated third-party observational data. However, AC statistics indicated agreement between the two methods was not exact, with mean ISP-Function and SDO values tending to not fall on the same point along the DBR scale for at least two of the targets. Multiple explanations for AC results are possible. For instance, agreement findings are likely reflective of the observed variability in data. ISP-Function Adult Attention and Peer Attention scores were the most variable across students, whereas the Escape/Avoidance and Access to Items/Activities were the least variable and consistently approximated zero. That the former targets tended to perfectly agree less with SDO while the latter demonstrated greater agreement is thus not surprising.

Overall, the extent to which such approximate but inexact observed agreement is acceptable is founded on the question of whether a 1-2-point difference might impact function-related decision making. It could be argued that such score differences would not be problematic as long as they are consistent across consequence targets, thereby not affecting the ranking of targets relative to each other. Consider a scenario where all four consequence targets are underestimated by one point. In this scenario, if one consequence score remains larger than the others, we would still conclude that the behavior is likely maintained by that former consequence, even if our estimates of their prevalence was not exact. On the other hand, if some consequence targets are underestimated while others are overestimated (or predicted exactly), the relative ranking of consequences would change and so would decisions regarding the likely function of behavior. As a result, additional research is necessary to evaluate the consequential validity of function-related decisions made in consideration of ISP-Function data. Should findings suggest ISP-Function-based decisions are consistent with SDO-based decisions, findings of inexact agreement between the two methods would be considered of less concern.

Limitations

Multiple limitations should be noted. One limitation was the lack of observational data for certain types of consequences. For example, there were few instances of students escaping or avoiding stimuli or gaining access to items or activities. Although this is typical of classrooms where teachers are commonly working to prevent such consequences from occurring in response to problem behavior, this lack of data makes it difficult to evaluate the performance of these ISP-Function targets. Results may look different in a context where a larger range of consequences are evidenced. Future research with more observational data points across a larger sample of students may help ameliorate this concern. Alternatively, future studies might incorporate a purposive sampling approach, wherein researchers actively recruit students from all function categories, with such categorization based upon initial functional assessments. Finally, researchers could examine the extent to which alternative definitions of these consequences influences estimates of their base rates across both methods.

Second, IOA data were only collected for 14% of observations. Future investigations should examine IOA for a larger percentage of observations, thereby verifying the performance of the criterion variable. Third, the diversity of the sample was relatively limited in terms of the racial/ethnic background of students. Specifically, the current sample included predominantly Caucasian and African American students from two regions of the country. As a result, future research should consider increasing student diversity from other geographic regions, as well as expanding the age range of students to include older students.

Implications for Research and Practice

Taken together, results demonstrate that pending additional research, the ISP-Function could be a highly efficient and low-inference method of collecting data on student behavior. When schools are shy of resources and personnel available to support the multitude of student behavioral needs present within a building, the ISP-Function moves past reliance on an outside observer to assess and monitor student behavioral concerns. Instead, teachers and other school staff working with the student are able to serve as the sole informant when monitoring student behavioral concerns and collecting data regarding the function of such concerns. In this manner, this measure permits efficient collection of data that inform timely intervention-related decisions for students. However, as with many assessment tools, future research is needed to delineate specific recommendations for use both within and across various assessment purposes, instrumentation, and procedures. Specifically, additional research is needed to examine the utility and acceptability of the ISP-Function within school settings as educators move towards multi-tiered models to support effective behavioral assessment and intervention frameworks.

References

Bruni, T. P., Drevon, D., Hixson, M., Wyse, R., Corcoran, S., & Fursa, S. (2017). The effect of functional behavior assessment on school-based interventions: A meta-analysis of single-case research. *Psychology in the Schools, 54,* 351-369. doi: 10.1002/pits.22007

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education and Treatment of Children*, *34*, 575-591. doi: 10.1353/etc.2011.0034

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., & Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48,* 219-246. doi: 10.1016/j.jsp.2010.02.001

Chafouleas, S. M., Kilgus, S. P., Riley-Tillman, T. C., Jaffery, R., & Harrison, S. (2012). Preliminary evaluation of various training components on accuracy of Direct Behavior Ratings. *Journal of School Psychology, 50,* 317-334. doi: 10.1016/j.jsp.2011.11.007

Chafouleas, S. M., McDougal, J. L., Riley-Tillman, T. C, Panahon, C. J., & Hilt, A. M. (2005). What do daily behavior report cards (DBRCs) measure? An initial comparison of DBRCs with direct observation for off-task behavior. *Psychology in the Schools, 42,* 669-676. doi: 10.1002/pits.20102

Cheney, D., Flower, A. L., & Templeton, T. (2008). Applying response to intervention metrics in the social domain for students at risk of developing emotional or behavioral disorders. *Journal of Special Education,* 42, 108-126. doi: 10.1177/0022466907313349

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention*, *34*, 201-213. doi: 10.1177/1534508409340390

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. doi: 10.1037//1040-3590.6.4.284

Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Dunlap, G., & Kern, L. (2018). Perspective on functional (behavioral) assessment. *Behavioral Disorders, 43,* 316-321. doi: 10.1177/0198742917746633

Durand, V. M., & Crimmins, D. B. (1988). Identifying the variables maintaining self-injurious behavior. *Journal of Autism and Developmental Disorders, 18,* 99-117. doi: 10.1007/bf02211821

Eckert, T. L., Martens, B. K., & DiGennaro, F. D. (2005). Describing antecedent-behavior-consequence relations using conditional probabilities and the general operant contingency space: A preliminary investigation. *School Psychology Review*, *34*, 520-528.

Filter, K. J., & Horner, R. H. (2009). Function-based academic interventions for problem behavior. *Education and Treatment of Children*, *32*, 1-19. doi: 10.1353/etc.0.0043

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders, 37,* 55-77. doi: 10.1177/019874291203700201

Ingram, K., Lewis-Palmer, T., & Sugai, G. (2005). Function-based intervention planning: Comparing the effectiveness of FBA function-based and non—function-based intervention plans. *Journal of Positive Behavior Interventions*, *7*(4), 224-236.doi: 10.1177/10983007050070040401

Iwata, B. A., & DeLeon, I. G. (1995). The functional analysis screening tool (FAST). *Unpublished manuscript, University of Florida*.

Iwata, B. A., DeLeon, I. G., & Roscoe, E. M. (2013). Reliability and validity of the functional analysis screening tool. *Journal of Applied Behavior Analysis, 46,* 271-284. doi:10.1002/jaba.31

Kilgus, S. P., Kazmerski, J. S., Taylor, C. N., & von der Embse, N. P. (2017). Use of Direct Behavior Rating to collect functional assessment data. *School Psychology Quarterly, 32,* 240-253. doi: 10.1037/spq0000156

Kilgus, S. P., Riley-Tillman, T. C., Stichter, J. P., Schoemann, A. M., & Bellesheim, K. (2015). Reliability of Direct Behavior Ratings–Social Competence (DBR-SC) data: How many ratings are necessary? *School Psychology Quarterly*, *31*, 431-442. doi: 10.1037/spq0000128

Klein, D. (2018). Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal, 18,* 871-901.

Klingbeil, D. A., Dart, E. H., & Schramm, A. L. (2018). A systematic review of function-modified check-in/check-out. *Journal of Positive Behavior Interventions*, 1-16. doi: 10.1177/1098300718778032

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. *What Works Clearinghouse*. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

LeBel, T. J., Kilgus, S. P., Briesch, A. M., & Chafouleas, S. (2010). The impact of training on the accuracy of teacher-completed direct behavior ratings (DBRs). *Journal of Positive Behavior Interventions*, *12*(1), 55-63. doi: 10.1177/1098300708325265

Lewis, T. J., Mitchell, B. S., Harvey, K., Green, A., & McKenzie, J. (2015). A comparison of functional behavioral assessment and functional analysis methodology among students with mild disabilities. *Behavioral Disorders*, *41*(1), 5-20. doi: 10.17988/0198-7429-41.1.5

Lloyd, B. P., Weaver, E. S., & Staubitz, J. L. (2017). Classroom-based strategies to incorporate hypothesis testing in functional behavior assessments. *Beyond Behavior, 26,* 48-56. doi:10.1177/1074295617711145

Martens, B. K., DiGennaro, F. D., Reed, D. D., Szczech, F. M., & Rosenthal, B. D. (2008). Contingency space analysis: An alternative method for identifying contingent relations from observational data. *Journal of Applied Behavior Analysis*, *41*(1), 69-81. doi: 10.1901/jaba.2008.41-69

McIntosh, K., Brown, J. A., & Brogmeier, C. J. (2008). Validity of functional behavior assessment within a response to intervention framework. *Assessment for Effective Intervention, 34,* 6-14. doi: 10.1177/1534508408314096

McIntosh, K., Campbell, A. L., Carter, D. R., & Dickey, C. (2009). Differential effects of a tier two behavior intervention based on function of problem behavior. *Journal of Positive Behavior Interventions*, *11*(2), 82-93. doi: 10.1177/1098300708319127

Newcomer, L. L., & Lewis, T. J. (2004). Functional behavioral assessment: An investigation of assessment reliability and effectiveness of function-based interventions. *Journal of Emotional and Behavioral Disorders*, *12*(3), 168-181. doi: 10.1177/10634266040120030401

O'Neill, R. E., Albin, R. W., Store, K., Horner, R. H., & Sprague, J. R. (2015). *Functional assessment and program development for problem behavior: A practical handbook* (3rd ed.). Stamford, CT: Cengage Learning.

Reinke, W. M., Herman, K. C., & Stormont, M. (2013). Classroom-level positive behavior supports in schools implementing SW-PBIS: Identifying areas for enhancement. *Journal of Positive Behavior Interventions*, *15*(1), 39-50. doi: 10.1177/1098300712459079

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A. M., & Glazer, A. D. (2008). Examining the agreement of Direct Behavior Ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions, 10,* 136-143. doi:10.1177/1098300707312542

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). Assessment (10th ed.). Boston: Houghton Mifflin.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428. doi: 10.1037//0033-2909.86.2.420

South, B. N. (2017). *Paraprofessionals' use of Direct Behavior Rating for collecting functional behavior assessment data: Agreement and acceptability* (Doctoral dissertation). Retrieved from ProQuest Dissertation & Theses A&I. (Order No. 10687657)

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*, 360-363.

Walker, V. L., Chung, Y., & Bonnet, L. K. (2018). Function-based intervention in inclusive school settings: A meta-analysis. *Journal of Positive Behavior Interventions, 20,* 203-216. doi: 10.1177/1098300717718350

Young, A., Andrews, C., Hayes, C., & Valdez, C. (2018). Should teachers learn how to formally assess behavior? Three educators' perspectives. *International Journal of Special Education, 33,* 416-426.

Zarcone, J. R., Rodgers, T. A., Iwata, B. A., Rourke, D. A., & Dorsey, M. F. (1991). Reliability analysis of the Motivation Assessment Scale: A failure to replicate. *Research in Developmental Disabilities*, *12*(4), 349-360. doi: 10.1016/0891-4222(91)90031-m

Table 1

Participant Demographic Characteristics

Variable and category	Teachers	Students
Grade (Student Only)		
Kindergarten		6
First Grade		9
Second Grade		6
Third Grade		7
Fourth Grade		4
Fifth Grade		2
Gender		
Female	33	9
Male	1	25
Race/Ethnicity		
White or Caucasian	28	18
Black or African American	2	15
Asian	0	1
Other	0	
Do not wish to answer	4	

Table 2

Descriptive Statistics

Scale	M	SD	Min	Max	Skewness	Kurtosis
ISP-DB	2.67	1.72	0.33	6.67	0.75	-0.11
ISP-AA	2.63	2.47	0.00	8.67	1.11	0.57
ISP-PA	1.77	2.11	0.00	8.00	1.65	2.48
ISP-EA	1.03	1.30	0.00	4.33	1.17	0.11
ISP-AIA	0.30	0.46	0.00	1.67	1.64	1.89
SDO-DB	2.88	1.41	0.67	6.67	0.85	0.48
SDO-AA	2.99	1.89	0.00	6.03	-0.09	-1.11
SDO-PA	2.75	2.22	0.00	9.03	0.91	0.59
SDO-EA	0.20	0.77	0.00	3.70	4.02	15.76
SDO-AIA	0.02	0.14	0.00	0.83	5.83	34.00

Note: ISP = Intervention Selection Profile–Function; SDO = Systematic direct observation; DB = Disruptive behavior; AA = Adult attention; PA = Peer attention; EA = Escape/avoidance; AIA = Access to Items/Activities.

Table 3

Temporal Reliability Coefficients and Minimum Number of Data Points Required for Acceptable ISP-Function Reliabilities

ISP-Function Target	ICC	Reliability	*r** = .70
Disruptive Behavior	0.22	0.46	8.18
Adult Attention	0.47	0.72	2.69
Peer Attention	0.50	0.75	2.35
Escape/Avoidance	0.14	0.32	14.95
Access to Items/Activities	0.12	0.29	17.50

Note: ICC = Intraclass correlation coefficients; *r** = minimum acceptable reliability level

Table 4

Correlations (ρ) between ISP-Function Ratings and SDO Scores

	Observation	ρ
Disruptive Behavior	Mean	.62**
	1	.73**
	2	.77**
	3	.49**
Adult Attention	Mean	.65**
	1	.53**
	2	.51**
	3	.54**
Peer Attention	Mean	.55**
	1	.53**
	2	.58**
	3	.65**
Escape/Avoidance	Mean	.22
	1	.30
	2	.29
	3	.30
Access to Items/Activities	Mean	-.15
	1	-.11
	2	-
	3	-

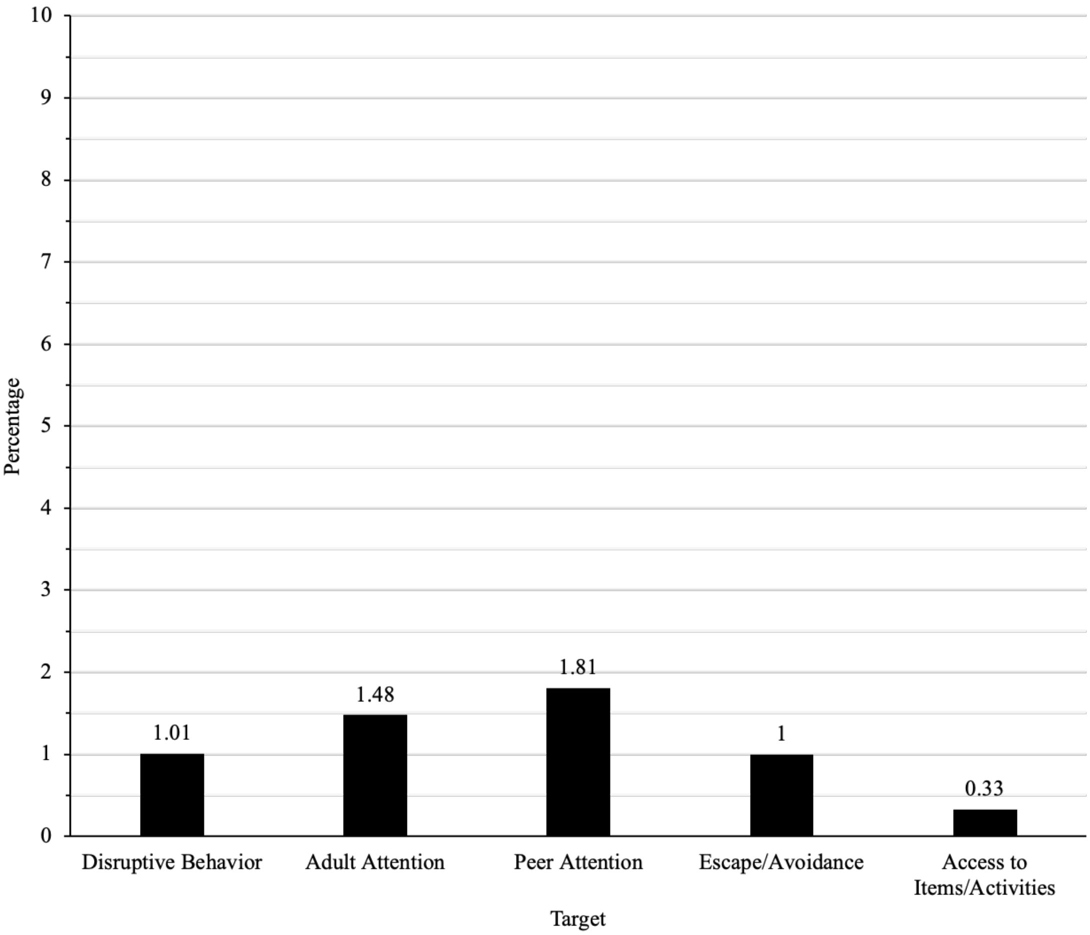

Figure 1. Mean difference scores across targets.