Citation: Matta, M., Volpe, R., Briesch, A.M. & Owens, J.S. (2020) Five Direct Behavior Rating Multi-Item Scales: Sensitivity to the Effects of Classroom Interventions. *Journal of School Psychology.*

Five Direct Behavior Rating Multi-Item Scales: Sensitivity to the Effects of Classroom Interventions

Michael Matta

University of Houston

Robert J. Volpe and Amy M. Briesch

Northeastern University

Julie Sarno Owens

Ohio University

Author Note:

Michael Matta, Department of Psychological, Health, and Learning Sciences, University of Houston; Robert J. Volpe, Department of Applied Psychology, Northeastern University; Amy M. Briesch, Department of Applied Psychology, Northeastern University.

This research was supported by a grant from the Institute of Education Sciences National Center for Special Education Research (R324A150071).

Correspondence should be addressed to Michael Matta, Department of Psychological, Health, and Learning Sciences, University of Houston, Houston, TX 77004.

E-mail: mmatta@uh.edu

Abstract

Direct Behavior Rating (DBR) is a tool designed for the assessment of behavioral changes over time. Unlike methods for summative evaluations, the development of progress monitoring tools requires evaluation of sensitivity to change. The present study aimed to evaluate this psychometric feature of five newly developed DBR Multi-Item Scales (DBR-MIS). Teachers identified students with behaviors interfering with their learning or the learning of others and implemented a Daily Report Card (DRC) intervention in the classroom settings for two months. The analyses were performed on 31 AB single case studies. Change metrics were calculated at an individual level by using $Tau\text{-}U_{\text{A vs. B + trend B}}$ and Hedges' g and at a scale-level by using Mixed Effect Meta-Analysis, Hierarchical Linear Models (HLMs), and Between-Case Standardized Mean Difference (BC-SMD). HLMs were estimated considering both fixed and random effects of intervention and linear trend within the intervention phase. The results supported sensitivity to change for three DBR-MIS (i.e., Academic Engagement, Organizational Skills, and Disruptive Behavior), and the relative magnitudes were consistent across the metrics. Sensitivity to change of DBR-MIS Interpersonal Skills received moderate support. Conversely, empirical evidence was not provided for sensitivity to change of DBR-MIS Oppositional Behavior. Particular emphasis was placed on the intervention trend in that responses to behavioral interventions might occur gradually or require consistency over time in order to be observed by raters. Implications for the use of the new DBR-MIS in the context of progress monitoring of social-emotional behaviors are discussed.

Keywords: Progress Monitoring, Direct Behavior Rating, Sensitivity to Change, Single Case Study Design, Tau-U, BC-SMD

Five Direct Behavior Rating Multi-Item Scales: Sensitivity to the Effects of Classroom Interventions

Schools increasingly have been using evidence-based multi-tiered systems of supports (MTSS) to address student academic and behavioral difficulties (Benner, Kutash, Nelson, & Fisher, 2013). With this shift to MTSS, the nature of behavioral assessment in schools has moved from primarily diagnostic or summative evaluation to early identification of problematic behaviors and recurring formative assessment. Within such frameworks, all students are assessed for social-emotional problems in order to identify discrepancies between current and expected levels of performance (i.e., universal screening). Targeted interventions are then implemented proactively with those students identified as at risk as opposed to waiting until significant issues arise. Students receiving intervention are monitored through formative behavioral assessment (i.e., progress monitoring) in order to ascertain whether interventions should be maintained, modified, or discontinued (Chafouleas, Riley-Tillman, & Sugai, 2007; Cook, Volpe, & Delport, 2013). As such, the success of MTSS relies on the availability of both evidence-based interventions and assessments.

Whereas extensive research has been conducted on evidence-based intervention programs, the body of literature on evidence-based assessment is comparatively small. This is especially true for progress monitoring tools of social-emotional constructs (e.g., Chafouleas, Volpe, Gresham, & Cook 2010; Dart, Arora, Collins, & Doll, 2019; Jensen-Doss et al., 2018). Existing tools are limited in several ways, including (a) a narrow range of social-emotional domains measured, (b) the predominant focus on problematic as opposed to positive behaviors, and (c) sparse evidence for psychometric adequacy.

Through an iterative process, a set of DBR measures was created to purposefully address the aforementioned limitations; we designed five scales for the assessment of key academic enablers (i.e., interpersonal skills, academic engagement, and organizational skills) and externalizing behaviors (i.e., disruptive and oppositional behavior). In previous studies, psychometric adequacy of the scales was investigated in terms of factor structure and reliability coefficients (Daniels, Briesch, Volpe, & Owens, 2019; Volpe, Chaffee, Yeung, & Briesch, 2020). The goal of the current study was to provide evidence for sensitivity to change by conducting a series of single case studies and integrating the results at the scale level.

<Direct Behavior Rating (DBR)>

DBR is a relatively new method of assessment that allows for a defensible, flexible, efficient, and repeatable formative evaluation of student social-emotional skills (Briesch, Chafouleas, & Riley-Tillman, 2016). DBR is defined by two core features: (a) behaviors are operationally defined, and (b) ratings are conducted immediately after a pre-determined interval (e.g., one activity block, one school day; Christ, Riley-Tillman, & Chafouleas, 2009). Practically, DBR combines the strengths of both systematic direct observation (SDO) and behavior rating scales. DBR is similar to SDO, in that rated behaviors are observable (and therefore designed to involve low levels of inference on the part of the rater), selected based on social importance (i.e., extent to which they are considered of great value by educators or detrimental for students), and evaluated in close temporal proximity to their occurrence. DBR is also similar to brief rating scales, in that ratings are based on the observations by key stakeholders (e.g., classroom teachers) and require little training to complete. Moreover, the involvement of raters who spend significant amounts of time with the student allows for the assessment of a broader range of

behaviors, including those with low base rates (i.e., low frequency; e.g., Daniels et al., 2019; Volpe et al., 2020).

The majority of studies investigating DBR have focused on single-item scales (DBR-SIS; Chafouleas, Riley-Tillman, & Christ, 2009) wherein an operational definition of a construct (e.g., academic engagement) is provided, and informants rate the presence of behaviors satisfying the definition on a continuous scale (e.g., 0 to 100%; Chafouleas et al., 2013; Kilgus, Chafouleas, Riley-Tillman, & Welsh, 2012; Kilgus, Riley-Tillman, Chafouleas, Christ, & Welsh, 2014). Numerous studies have provided support for the dependability, efficiency, and sensitivity to change of DBR-SIS measuring academic engagement and disruptive behavior (e.g., Chafouleas, Sanetti, Kilgus, & Maggin, 2012; Miller, Crovello, & Chafouleas, 2017; von der Embse, Scott, & Kilgus, 2015).

Over the past several years, there has been increased interest in DBR comprised of multiple items. DBR Multi-Item Scales (DBR-MIS; Volpe & Briesch, 2012) resemble brief rating scales in that informants are asked to rate several items (typically three to five) that are summed to generate a composite score for the construct of interest (e.g., academic engagement). Among the advantages of DBR-MIS are that it (a) may require fewer assessment occasions to generate data with sufficient dependability comparted to DBR-SIS (Volpe & Briesch, 2012), (b) allows for customization of measurement based on the presenting problems of individual students (Volpe & Briesch, 2015, 2016; Volpe & Gadow, 2010), and (c) affords the ability to measure specific behaviors (item-level data) in addition to the broader assessment of the construct to which the items serve as indicators (scale-level data). Consequently, DBR-MIS may offer enhancements in efficiency in addition to providing enhancements to the granularity of

assessment that may aid in decision-making (Volpe & Briesch, 2015; Volpe, McConaughy, & Hintze, 2009).

Although the extant research concerning DBR-MIS is encouraging, unfortunately, few studies have examined the degree to which these measures are sensitive to the effects of interventions (Daniels, Volpe, Briesch, & Gadow, 2017; Hustus, Owens, Volpe, Briesch, & Daniels, 2020; Volpe & Gadow, 2010). Moreover, the majority of these investigations were focused on general education students rather than students at risk for the development of behavioral problems (and for whom interventions are of primary importance) (e.g., von der Embse et al., 2015; Fabiano, Pyle, Kelty, & Parham, 2017). The assessment of non-academic student behaviors, at least for progress monitoring purposes, is likely to be limited to those students exhibiting problematic behaviors in applied settings (e.g., Tier 2 or 3 of MTSS). In our examinations of sensitivity to change, we carefully designed participant selection procedures to ensure that study participants represented the population of students who would typically be assessed within the proposed formative evaluation system.

<Sensitivity to Change of DBR-MIS>

Historically, if a measure was to demonstrate evidence of reliability and validity, it would be considered suitable and ready for applied use (Stratford et al., 1996). However, Guyatt, Walter, and Norman (1987) argued that "what we would really like to know about an evaluative instrument is the likelihood of detecting a clinically important treatment effect, even if that effect is small" (p. 174). This psychometric characteristic is known as sensitivity to change (or "treatment sensitivity") and has been investigated by using either between-group or single-case designs.

Within the context of between-group studies, sensitivity to change is calculated by investigating the presence of significant differences between two or more groups, one of which is generally not exposed to the treatment and serves as a control group. Within the context of single-case studies, sensitivity to change is calculated by comparing student behaviors during the baseline phase with the data collected during or after the implementation of an intervention. This procedure can be repeated across multiple participants or within the same participants two or three times in order to improve the experimental control and the generalizability of the results (Byiers, Reichle, & Symons, 2012). Numerous change metrics – often named "effect sizes" (Carter, 2013) – have been suggested to investigate sensitivity to change of new assessment methods, each with strengths and weaknesses (Brossart, Laird, & Armstrong, 2018; Gresham, 2005; Parker, Vannest, Davis, & Sauber, 2011; Pustejovsky, Hedges, & Shadish, 2014). Such change metrics can be divided into two categories, namely non-parametric statistics (also known as "non-overlap measures") and parametric statistics. Non-parametric statistics (e.g., Percent of Data Exceeding the Median of Baseline [PEM], Nonoverlap of All Pairs [NAP], *Tau-U* coefficients) quantify the degree of non-overlap between baseline and intervention phases and are not based on assumptions regarding the distribution of the dependent variable. Parametric statistics (e.g., Standardized Mean Difference [SMD], Hedges' *g*, Between-Case Standardized Mean Difference **[**BC-SMD]) rely on the assumption that the outcomes follow known distributions (e.g., Gaussian, Poisson, etc.), are associated with sampling variances, and are less sensitive to procedural characteristics of the study design (Pustejovsky, 2019), such as the length of the phases of the observation session. The interpretation of the magnitude of the change metrics is typically consistent with Cohen's (1988) considerations (i.e., null effects for values

lower than 0.20, small effects between 0.20 and 0.50, moderate between 0.50 and 0.80, and values greater than 0.80 indicate large or very large effects).

Finally, although relatively new in the field of education, Hierarchical Linear Modeling (HLM), also known as Multi-Level Modeling (MLM), has been used as a methodological bridge between single-case and between-group designs in that it affords the ability to account for both within-individual and across-individual variance across measurement occasions. Further, the interaction between phase and time (i.e., linear or polynomial trend) can be added to models to investigate change over time both within the baseline and the intervention phases (Pustejovsky et al., 2014; Van den Noortgate & Onghena, 2008).

Four studies to date have provided evidence for the sensitivity to change of DBR-MIS, with three of these studies providing such evidence in response to stimulant medication. Volpe, Gadow, Blom-Hoffman, and Feinberg (2009) first examined the psychometric characteristics of 9-item and 4-item scales measuring inattentive and hyperactive-impulsive behaviors in the context of a school-based medication titration study involving a no-treatment condition and three doses of methylphenidate. Results of repeated measures MANOVA (RM-ANOVA) indicated that each scale demonstrated sensitivity to the effects of stimulant medication. Subsequently, Volpe and Gadow (2010) explored the properties of two sets of 3-item scales to detect small changes in students' behaviors when receiving three doses of medications for inattentive and hyperactive behaviors, aggression, and conflict with peers over the course of two weeks. Similar to the earlier study, an RM-MANOVA indicated that the scales were sensitive to the effects of stimulant treatment. However, in this later study Volpe and colleagues (2010) found that, for certain constructs, increased sensitivity to change was demonstrated when the content of measures was customized for each student based on their unique constellation of

symptomatology. Daniels and colleagues (2017) re-examined the scale measuring peer conflict by selecting a pool of six items; support for adequate sensitivity to change was confirmed by four change metrics (i.e., absolute change, effect size, percentage change from the baseline, and reliable change index) that compared scores from three days of baseline and three days of treatment.

Although the results of these studies have demonstrated sensitivity of DBR-MIS to the effects of intervention, all evaluated the effects of stimulant medication, which are known to be associated with large effects compared to non-treatment controls (e.g., Prasad et al., 2013). To date, only one study has assessed sensitivity to change in the context of a typical classroom-based intervention. Our team (Hustus et al., 2020) conducted a pilot investigation of the sensitivity to change of four DBR-MIS assessing academic engagement, organization skills, disruptive behavior, and oppositional behavior. Behaviors of five kindergarten to fourth-grade students were rated on a daily basis while teachers implemented daily report card (DRC) interventions over the course of two months. Non-overlapping metrics (i.e., Tau_{novlap} and *Tau-U*) supported sensitivity to change of DBRs measuring disruptive and engagement; results for organizational skills were less convincing and for oppositional were inconclusive. Although promising, such results provided preliminary evidence because (a) the sample size was composed of five students only; (b) participants were rated each on different scales, hence preventing the calculation of change metrics at the DBR level; and (c) the presence of missing data was not adequately considered in the data analyses.

<Purpose of Study>

Given limited evidence to date, the goal of the current study was to assess sensitivity to change for five DBR-MIS in the context of individualized classroom-based interventions. We

expected that the five DBR-MIS would exhibit sufficient sensitivity to change. That is, we hypothesized that the results would show significant behavioral changes (i.e., magnitude of the overall effects equal to or higher than .20; Cohen, 1988) and would follow the predicted directions (i.e., improvement for scales measuring academic enablers and reduction for externalizing behaviors).

This study enriches the knowledge base on the sensitivity to change of DBR-MIS by expanding the preliminary findings of Hustus and colleagues (2020) in three main directions. First, in addition to sensitivity to change at individual-level (as conducted in Hustus et al., 2020), analyses also were conducted at the scale-level in order to provide more robust and generalizable estimates. That is, whereas students were evaluated on scales measuring different behavioral domains in Hustus et al. (2020), within the current study, we conducted analyses on multiple students whose behavior was evaluated on the same DBR-MIS, thus providing evidence of sensitivity to change beyond individual levels. Second, we introduced a new DBR-MIS measuring interpersonal skills; the evaluation of these skills is of great importance for students because interpersonal deficits are associated with a wide variety of disabilities interfering with learning processes, hence are often the target of classroom interventions (Elliott, Malecki, & Demaray, 2001). Third, we investigated sensitivity to change via HLM, recently adapted to single case studies (Valentine, Tanner-Smith, Pustejovsky, & Lau, 2016), which allows for the estimation of both fixed and random effects of the predictors and for testing both the shifts in the outcome mean and the presence of linear trends within one or both phase.

<Method>

<Participants>

Data were collected during the 2016-2017 and 2017-2018 school years. Participants included19 kindergarten through fourth-grade general and special education teachers recruited across two sites (i.e., rural Midwest and urban Northeast). All teachers were Non-Hispanic Caucasian, the majority were female (95%), and their ages ranged from 23 to 53 (M = 39.41). Each teacher referred one student (see details below); however, one student was referred by both a general and special education teacher. Therefore, 18 students were involved in the study (data from six of these students were analyzed in the pilot study, Hustus et al., 2020). Most of the students were Non-Hispanic Caucasian (N = 15), male (N = 16), and their age ranged from 6 to 10 (M = 7.72). The study was approved by the Institutional Review Board of both universities and by school district administrators.

<Measures>

Integrated Screening and Intervention System Teacher Report Form (ITRF; Volpe & Fabiano, 2013). The ITRF is a 43-item screening form, which focuses on observable and malleable problem behaviors (e.g., disrupts others, moves around the room) rather than diagnostic symptoms. Each item is rated on a 3-point scale, ranging from 1 (Slight Concern) to 3 (Strong Concern). The ITRF allows teachers to rate simultaneously up to five students who exhibit concerning behaviors. A total score is computed by summing item ratings; students with a total score of 30 or higher are likely at risk for demonstrating problematic behaviors (Daniels et al., 2017) and would benefit from targeted behavioral intervention. The ITRF has shown high internal consistency (α = .97), strong temporal stability from two weeks to one month (r = .84), and evidence for convergent validity (r > .81) with other behavioral teacher report measures (Daniels et al., 2014).

DRC implementation data. Teachers were asked to give students feedback when a DRC

rule violation occurred (e.g., *Carlos, that's an interruption*) and to document the student's performance toward each goal (e.g., tallies for interruptions, percent of work complete) each day. Teachers were asked to submit these data (either into a website that produced graphs of student performance or to the project consultant). These data were used to match DRC implementation days to DBR-MIS completion days.

Treatment Integrity Form (TIF; Volpe & Fabiano, 2013). The TIF is a 9-item checklist that assesses DRC implementation behaviors. Each item is rated on a dichotomous scale based on whether the observer (e.g., research staff) considers that the teacher adopted appropriate procedures at least half of the time (e.g., *feedback on DRC targets was provided, reward was provided if the child attained DRC goal*). The checklist required the observer to indicate the number of DRC goal violations by the teacher during the time period of the observation. The information was recorded for all goals on the student's DRC. The total score is calculated as the proportion of items performed by the teacher divided by the total number of applicable items.

Direct Behavior Rating Multi-Item Scale (DBR-MIS). Five DBR-MIS were used to assess academic enablers and problem behaviors for this study. The academic enabler scales included the measurement of interpersonal skills (e.g., *respectful; cooperates*), academic engagement (e.g., *works independently; on task)*, and organization skills (e.g., *keep track of assignments and materials; completes assignments*), whereas the problem behavior scales measured disruptive (e.g., *talks to classmates when inappropriate*; *calls out*) and oppositional behavior (e.g., *disrespectful; uncooperative*). Each DBR-MIS consisted of five items, and teachers were asked to rate on a 7-point scale either (a) how often the academic enabling behaviors were exhibited during the day (i.e., *Never* to *Almost Always*) or (b) how much of a problem the behaviors were perceived to be (i.e., *Not a Problem* to *A Serious Problem*).

Exploratory factor analyses demonstrated unidimensionality for each DBR-MIS and strong loadings for the items on the corresponding latent factor (ranging from .75 to .92). Evidence of internal consistency was demonstrated by alpha coefficients higher than .90 across the five DBR-MIS (Daniels et al., 2019; Volpe et al., 2020).

<Study Design and Procedures>

A series of single-subject AB designs were used to assess sensitivity to change of the five DBR-MIS in the context of an eight-week DRC intervention. The DRC was chosen as the intervention method because its effectiveness has been widely tested and documented within the classroom environment (Vannest, Davis, Davis, Mason, & Burke, 2010). Teachers consider DRC to be an appropriate tool because it can be adapted to address a wide variety of behaviors, it is feasible, and its scores provide immediate information about whether the intervention is successful (Fabiano et al., 2017; Owens et al., 2012).

Single-subject AB designs are frequently used to establish evidence of correlational relationships between interventions and outcomes and to provide support that observed changes are associated with an implemented intervention rather than a function of the passage of time (Chafouleas et al., 2012; Riley-Tillman & Burns, 2009). Although the lack of randomization prevents from the possibility of making causal claims, “inferences regarding the correlation between changes in independent and dependent variables” (Kilgus, Riley-Tillman, & Kratochwill, 2016, p. 481) can be derived from AB designs in the context of lower stake decisions (e.g., evaluating Tier 2 intervention's effectiveness) and in the presence of standardized, evidence-based interventions (e.g., DRC). Additionally, we enhanced the defensibility of the current study design by attempting to replicate similar effects across participants and across two sites (Kilgus et al., 2016; Kratochwill & Levin, 2010).

Participating teachers were asked to sign a consent form and rate up to five students on the ITRF who demonstrated behaviors that interfered with their learning or the learning of others and would benefit from a behavioral intervention. Students who obtained a score of 30 or higher (Daniels et al., 2017) were considered eligible for study participation; when two or more students were eligible, teachers were asked to select the student who had obtained the highest score. Teachers then sent a copy of the study description and consent forms to the eligible students' parents. Interested parents were given the opportunity to ask any questions of the Principal Investigator before signing the form. Once they provided written consent, a research assistant met with the student to provide a friendly explanation of the study and obtain their assent. If parents declined to allow their child to participate, the teacher selected the second-ranked student on the ITRF and followed the same procedures.

Once the three consent forms were collected (i.e., teacher, parent, and student), a research assistant met with the teachers and conducted a semi-structured interview (Target Behavior Interview – TBI; available at http://oucirs.org/daily-report-card website) to obtain information about their perspective on the selected student's strengths and problematic behaviors. The overall goal of the interview was to identify two to four behaviors that would be suitable targets for the DRC intervention. Research assistants placed particular emphasis on the ITRF items that teachers rated with a score of 3 (i.e., *Strong Concern*). The interview allowed for the identification of time frames during which behaviors were most likely to occur (e.g., all day, in one specific class).

Research assistants and the teacher selected one or two DBR-MIS for each student corresponding to the DRC targeted behaviors identified during the TBI, based on a grid developed by the authors. Prior to starting the study, six members involved in the research

identified the two DBR-MIS that best matched each ITRF item (grid available upon request from the corresponding author). The agreement among the team members was generally high, meaning that all six raters identified the same two DBR. For instance, the research team agreed that the ITRF item "The student does not complete class-work on time" most closely corresponded with the behaviors evaluated by DBR Academic Engagement and DBR Organizational Skills. If, during the interview with the teacher, the completion of class-work was of major concern, the team implemented the appropriate DRC goal (e.g., *complete assigned work on time*) and adopted one or both aligned DBRs to track progress over time. In cases in which consensus could not be reached unanimously, matches were considered appropriate when at least three out of the six members agreed.

Once DBR-MIS were selected, research assistants provided a start date to the teachers. Prior to implementing the DRC intervention, each teacher received a brief training from a research assistant, including how to introduce the DRC to, and review progress with, the student and procedures to track behaviors on the DRC. Depending on the teacher's experience and background knowledge, the training lasted between 5-15 minutes. Additionally, teachers received supplemental materials describing DRC development and procedures and recommendations for implementation (e.g., how to provide positive reinforcement and remind students about their DRC goals and progress). They were instructed to complete the two DBR-MIS using an online platform at the end of every day for the baseline period. Per What Works Clearinghouse (WWC) guidelines (Kratochwill et al., 2010), at least three data points were collected during the baseline phase or until a stable pattern of responding was observed. Then, the DRC intervention was implemented, and the DBR-MIS were tracked daily for up to eight weeks.

Research assistants checked in with teachers on a weekly basis, assisted teachers in making decisions regarding possible adjustments to the DRC, and provided support via email as needed. Research assistants also completed classroom observations for the duration of DRC interventions in order to evaluate teacher implementation integrity. After each observation, the observers completed the TIF, indicating adherence to its DRC implementation behaviors. All the treatment behaviors were observed and contribute to the evaluation of treatment integrity. Finally, the evaluation of sensitivity to change of each DBR-MIS relied on a different number of students because the overall sample was a consecutively-recruited convenience sample. That is, although we tried to recruit an equal number of students presenting with each behavioral concern or deficit, we were not always able to do so. However, a minimum of three cases was required for each scale to be included in the study; this decision was made to (a) be consistent with the literature on single-case studies wherein authors are generally asked to provide evidence of the effect from three different participants (Kratochwill et al., 2010), (b) avoid highly unstable overall estimates, and (c) model the effects of treatment and trend on the outcome (Valentine et al., 2016).

<Data Analysis>

First, the presence of missing data was determined at the individual level by matching the dates of DRC implementation data (the presence of daily data on student target behaviors) to DBR ratings. Two students were excluded from the analyses because the percentage of missing data was greater than 50%. Then, we tested whether the ratings were missing at random. We created a dichotomous variable (0 = missing rating, 1 = no missing rating) and run a set of logistic regressions where intervention and linear trend were entered as independent variables. None of the regression coefficients was associated with significant effects. Therefore, we

concluded that there was no relationship between the effectiveness of DRC intervention and teachers missing an opportunity to rate a student's behavior on the DBR-MIS. We conducted multiple imputations for time series by chained equations and created 30 completed data sets using the R package Amelia (Honaker, King, & Blackwell, 2011). The number of imputations was established according to the average percentage of missing data and simulation studies (Graham, Olchowski, & Gilreath, 2007; Royston & White, 2011). The parametric and non-parametric estimates obtained from each imputed data set were pooled by using Rubin's formulas (Enders, 2010).

Second, a treatment integrity score was calculated for each session. Then, these scores were weighted by the duration of the observation to obtain a single estimate at the teacher level, which, in turn, was weighted by the number of sessions conducted per student in order to measure treatment integrity for each DBR.

Third, the magnitude of sensitivity to change of the five DBR-MIS was analyzed both within and between cases in two stages. Following the WWC recommendations (Kratochwill et al., 2010), we selected more than one estimate for each approach given that there has been no consensus regarding the most appropriate measures for single-case studies. The selection process of the change metrics followed four steps. First, the metrics were grouped into two classes based on whether they were parametric or non-parametric (e.g., Parker et al., 2011; Pustejovsky, 2019). Second, change metrics with severe procedural limitations were excluded from consideration; for instance, the number of baseline and treatment sessions greatly affects the calculation of Percent of All Non-overlapping Data (PAND), and the presence of one 0 in the baseline session leads the calculation of Percentage of Zero Data (PZD) to 0. Third, other change metrics were ruled out because of their assumptions; for example, log-response ratios (Pustejovsky, 2018) require true

0s and thus are not consistent with at least some of the behavior rating scales. Finally, we selected change metrics among those more recently developed that would allow us to capture effects associated with shifts both in the outcome level and trend. We describe the selected metrics below.

Within-case effect sizes. Firstly, both nonparametric measures and parametric effect sizes were calculated within each case. Hedges' *g* was assessed by using the SingleCaseES R package (Pustejovski & Swan, 2017), and $Tau\text{-}U_{\text{A vs. B + trend B}}$ was calculated by using the R code adapted from Parker and colleagues' original paper by Tarlow (2017, March). In addition, data of single case studies were plotted with ggplot2 R package (Wickham, 2016). Appendix A includes the individual graphs grouped by DBR-MIS.

Within Case-Standardized Mean Difference (*WC-SMD*; Busk & Serlin, 1992; Hedges, 1981). Analogous to *SMD* used in between-group intervention studies, Hedges' *g* has been adapted to single-case designs by correcting the estimate for small samples. Psychometric assumptions for the underlying model include lack of significant trends, intra-individual residuals normally distributed around phase means, and a similar effect of the intervention across cases (Shadish, Hedges, Horner, & Odom, 2015).

When applied within-case, this measure is calculated as the difference between the mean of the intervention phase and the baseline phase, divided by the standard deviation of the baseline phase, which is considered constant across the two phases. Despite the similarities in the Hedges' *g* formula between single case and between-groups studies, the sum of the means is scaled respectively by intra-individual variability only and both intra- and inter-individual variability in the outcome. Therefore, the two methods result in different scales, and their values

are not directly comparable (Pustejovsky & Ferron, 2017; Van den Noortgate & Onghena, 2008). Hedges' *g* has been interpreted following the rule of thumb applied to Cohen's *d* (Cohen, 1988).

Tau-U (Brossart et al., 2018; Parker et al., 2011). *Tau-U* is a family of rank correlation measures that examines treatment effects on both between-phase overlap and within-phase trends in the same τ metric. The combination of the three regions within the difference matrix allows for the calculation of four *Tau-U* coefficients that may be distinguished with subscripts: (a) $Tau\text{-}U_{\text{A vs. B}}$ compares each pair of data points between the two phases; (b) $Tau\text{-}U_{\text{A vs. B – trend A}}$ and $Tau\text{-}U_{\text{A vs. B + trend B}}$ consider the improvement between the two phases and allows for the inclusion of the baseline or the intervention trend, respectively; and (c) $Tau\text{-}U_{\text{A vs. B – trend A + trend B}}$ represents the percentage of nonoverlapping data and controls for both baseline and intervention trends.

$Tau\text{-}U_{\text{A vs. B + trend B}}$ coefficient was selected as the metric for the calculation of sensitivity to change because this coefficient has the advantage of capturing either positive or negative deviations in the intervention phase that can be important indicators of improvements, especially when changes are expected to be gradual and over time. In addition, we assumed that no linear trend affected the baseline across the cases (see Appendix A). It is worth noting that the consideration of additional variance within the intervention phase tends to reduce the magnitude of *Tau-U* and improve the *p*-value significance level.

Between-case effect size estimates. Secondly, scale-level estimates were obtained. Mixed Effect Meta-Analysis was performed by using metafor R package (Viechtbauer, 2010), Two-Level Hierarchical Linear Model by using nlme R package (Pinheiro, Bates, DebRoy, Sarkar & R Core Team, 2019), and *BC-SMD* by using the scdhlm R package (Pustejovski, 2016).

Mixed Effect Meta-Analysis (Schwarzer, Carpenter, & Rücker, 2015). Fixed effects and heterogeneity measures were calculated to combine nested data from students whose behavioral changes were evaluated on the same scale. *Tau-U*$_{\text{A vs. B + trend B}}$ estimates for individual cases represented component studies and were averaged by using the "inverse variance method" and considering the level of precision. Fixed effect estimates were consistent with preliminary considerations about each scale. Heterogeneity was measured by between-study variance indices, such as tau^2, H and I^2, and tested using the Cochran's Q statistic and the corresponding p-value (Higgins & Green, 2008; Rücker, Schwarzer, Carpenter, & Schumacher, 2008). I^2 is a scaled version of H varying between 0 and 1, and values of I^2 near zero and smaller values of Q suggest that the estimate associated with the outcome is comparable across cases (Borenstein, Higgins, Hedges, & Rothstein, 2017).

Two-level hierarchical linear models (Baek, Petit-Bois, Van den Noortgate, Beretvas, & Ferron, 2016). HLMs represent an extension of linear regression models wherein outcomes are estimated primarily through the shift in level and trend. This approach is consistent with visual analysis; it is particularly appropriate when a study aims to evaluate the magnitude of the effect at specific time points and allows for the measurement of how intervention effects change over time and across either cases or studies and the contribution of significant moderators (Shadish, Kyse, & Rindskopf, 2013).

Such models can handle recurring issues involved in interrupted time trend series, such as significant associations in errors attributable to autocorrelation (also known as serial correlation), heterogeneity across variances, distribution patterns (either linear or nonlinear), and count outcomes data. Simulation studies have shown that fixed effects are robust to violations of normality assumptions (Moeyaert, Ugille, Ferron, Beretvas, & Van Den Noortgate, 2016).

In the present study, we modeled a two-level hierarchical structure where measurement occasions (i.e., first-level units) were nested within cases (i.e., second-level units). This approach allowed us to estimate intervention effects targeting similar behaviors across students.

Four sets of predictors were considered in relation to the DBR outcomes: (1) fixed and varying intercepts and fixed intervention effect, (2) fixed and varying intercepts and intervention effect, (3) fixed and varying intercepts and intervention effect and fixed linear trend, and (4) fixed and varying intercepts, intervention and linear trend effect. The equation formally defining the models was:

$$Y = \beta_0 + \beta_1 \, Intervention + \beta_2 \, Linear \, Trend + residual$$

where β_0 represents the target behavior prior to introducing the intervention, β_1 defines the immediate behavioral change in the outcome level after implementing the intervention, and β_2 represents additional changes in the outcome per session within the intervention phase. As for the residuals, we assumed first-order autoregressive covariance structure, meaning that the relationship between variances changes in a systematic way. This is a common assumption for repeated-measures data in that the correlation is highest at adjacent time points. The four regression models derived from the general equation were specified as follow:

Model 1: $\beta_0 = \gamma_0 + \eta_0; \quad \beta_1 = \gamma_1; \quad \beta_2 = 0$

Model 2: $\beta_0 = \gamma_0 + \eta_0; \quad \beta_1 = \gamma_1 + \eta_1; \quad \beta_2 = 0$

Model 3: $\beta_0 = \gamma_0 + \eta_0; \quad \beta_1 = \gamma_1 + \eta_1; \quad \beta_2 = \gamma_2$

Model 4: $\beta_0 = \gamma_0 + \eta_0; \quad \beta_1 = \gamma_1 + \eta_1; \quad \beta_2 = \gamma_2 + \eta_2$

where γ represents the grand mean of the effects, and η is assumed to follow a normal distribution with a mean of zero e variance τ^2. The four models rely on different assumptions: (a) Model 1 assumes stable behavioral outcome (i.e., lacking trend) prior to the intervention and

that the intervention is associated with a change in the outcome level which is constant across cases, (b) Model 2 relaxes the constraint of equal intervention effects across cases, (c) Model 3 includes the interaction between intervention and time and assumes that such incremental effects are constant across cases, and (d) Model 4 relaxes the constraint of equal interaction across cases.

The Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were used to select between the four alternative models. Such information criteria allow for the comparison of which subset of predictors best explains the dependent variable (Singer & Willett, 2003). Because there is no established procedure to pool the results for AIC and BIC, we computed their means and 95% confidence intervals for each model, as suggested by Enders (2010). Models with smaller values indicate a better fit. We adopted standard values for model selection suggested by Burnham and Anderson (2004) for AIC and by Raftery (1995) for BIC.

Finally, we calculated the weighted mean of slope reliability across cases for each DBR-MIS. Reliability of slope refers to the proportion of true slope variance to total slope variance and is an indicator of how well the data accurately represent student behavioral outcomes. Reliability of slope is important to formative assessment because strong coefficients are evidence that data are representative of student behavior rather than due to systematic or random sources of error (Christ, 2006). Reliability coefficients of .80 or greater are generally desirable. Coefficients ranging between .60 and .80 may be acceptable in the contexts of low stake decisions or under specific circumstances, whereas coefficients less than .60 are generally inadequate for supporting reliability of slope (Hintze & Marcotte, 2010).

Between-Case Standardized Mean Difference (*BC-SMD*; Pustejovsky et al., 2014). *BC-SMD* allows for the calculation of effect sizes for single case designs. When multiple observations are available from at least three participants, a hierarchical model with Restricted Maximum Likelihood (REML) procedure is applied in order to describe the correlational relationship between the intervention and the outcome for each individual and how the pattern of data changes across cases in the study.

BC-SMD has two important advantages. First, fixed and random effects of baseline, treatment, and trends contribute to the estimation of the effect size. Second, the coefficients are expressed in the *SMD*-metric that is commonly used by scholars who conduct studies on groups, increasing the likelihood for single case studies to be considered as evidence when evidence-based interventions are reviewed. The downside of *BC-SMD* is that it requires a minimum of three individuals; this may be appropriate to answer basic questions (such as about the effectiveness and the magnitude of interventions), but larger samples are needed for more sophisticated conclusions (such as the presence of covariates). The formal equation for the calculation of *BC-SMD* is:

$$\delta_{AB} = \frac{E[Y_B(A)] - E[Y_B(n)]}{\sqrt{Var\,[Y_B(n)]}}$$

where the numerator represents the difference between the average outcome if intervention is implemented after time *A* and the average outcome if intervention is not implemented (i.e., after time *n*), scaled by the square root of the outcome variance if the intervention is not implemented. The four effect sizes derived from the general equation were specified as follow:

Model 1 and 2: $E[Y_B(A)] = \gamma_1$ $\quad Var\,[Y_B(n)] = \tau_0^2 + \sigma^2$

Model 3 and 4: $E[Y_B(A)] = \gamma_1 + \gamma_3(B - A)$ $\quad Var\,[Y_B(n)] = \tau_0^2 + \sigma^2$

It is worth noting that the inclusion of random effects in the computation of *BC-SMD* does not change its formula; in fact, its computation involves the same parameters used for the fixed effects. The difference between the two effect sizes relies on how the parameters are calculated via HLM.

In addition, $B - A$ corresponds to the difference between the outcomes at a follow-up time and the session immediately following the intervention implementation; therefore, the selection of such parameters will greatly influence the magnitude of *BC-SMD*. In fact, effect sizes are dependent on the choice of a hypothetical treatment initiation time (A) and follow-up time (B) across cases. In the calculation of the effect size for Model 3, $B - A$ was set at the tenth intervention session after the intervention was implemented for the last student. Note that for some cases – typically the first students receiving the intervention – the selection of B involved substantial extrapolation past the observed intervention data. Finally, the denominator of Models 3 and 4 was equal to Models 1 and 2 because the predictors were centered at 0.

<Results>

Based on ITRF scores and teachers' interviews, (a) four students received a DRC intervention for interpersonal skills, academic engagement, or oppositional behaviors; (b) eight students received a DRC intervention for organizational skills; and (c) seven students received a DRC intervention for disruptive behaviors. On average, DRC data were collected for 48.81 days (SD = 12.08), ranging from 27 to 66. The rate of completion of DBR-MIS was approximately 87% during the baseline phase (Mdn = 9 occasions, range: 6–31) and approximately 81% during the intervention phase (Mdn = 39 occasions, range: 21–53). Descriptive statistics of each phase duration and completion rate by DRC domain are reported in Table 1.

<Treatment Integrity>

Treatment integrity data calculated by DRC domain are reported in Table 2. On average, treatment integrity observations were about 43 minutes long (*SD* = 13.46), and each teacher was observed for 6.47 occasions (*SD* = 2.65, range: 3–12). The overall teachers' adherence rate was 66%; however, there were noticeable differences across the domains. The highest adherence rates were noted for those DRC interventions targeting engagement and interpersonal skills (*M* = 75% and 74%, respectively), followed by disruptive behaviors and organizational skills (*M* = 69% and 67%, respectively). Finally, the lowest percentage of treatment integrity was noted for those DRC interventions targeting oppositional behaviors (*M* = 53%).

<Descriptive Statistics>

Weighted means, medians, and standard deviations were calculated by phase for the five DBR-MIS. The values showed trivial differences before and after performing multiple imputations (see Table 3), which means that the imputation of missing data did not change the distribution of the original DBR-MIS scores. Hence, all statistical analyses were conducted on the imputed data sets, and the results were then pooled together.

Overall, DBR-MIS mean and median scores demonstrated changes in the expected direction, with the median values associated with a larger difference because of its robustness to outliers. This result provided initial support for sensitivity to change of the scales in the context of an intervention. The DBR-MIS assessing academic enablers (i.e., Interpersonal Skills, Engagement, and Organizational Skills), as well as the DBR-MIS Disruptive Behavior, demonstrated notable changes, whereas the DBR-MIS Oppositional Behavior resulted in a small difference between the baseline and the intervention scores.

Variability increased between the two phases for the five DBR-MIS as well. Minimum and maximum scores were calculated within each phase. Table 4 shows that the average of

minimum scores was similar across the two phases, whereas the average of maximum scores displayed large differences and followed the expected direction of the intervention. Consistent with the previous results, the DBR-MIS Oppositional Behavior was not associated with notable differences either on minimum or on maximum scores.

<Within-Case Effect Size Estimate>

Pooled *Tau-U*$_{\text{A vs. B + trend B}}$ and Hedges' *g* were calculated as within-case effect sizes (see Table 5). Standard errors and confidence intervals were reported to describe the accuracy and precision of the estimates.

DBR-MIS Interpersonal Skills. Sensitivity to change of the DBR-MIS Interpersonal Skills was assessed by examining the changes in scores of four students who needed to improve their skills when interacting with others. In two cases, *Tau-U* coefficient and Hedges' *g* provided fair evidence for DBR responsiveness to the DRC intervention. For Students 2 and 4, *Tau-U* coefficients were associated with medium changes (i.e., *Tau-U* = 0.41 and 0.36, respectively), and Hedges' *g* showed from large to very large effects (i.e., *g* = 1.55 and 0.70). However, *Tau-U* was approximately null in the other two cases (Students 1 and 3, with the effect for Student 3 noted to be in a direction contrary to expectations), whereas Hedges' *g* was significant and very large for Student 1 (i.e., *g* = 2.55).

DBR-MIS Academic Engagement. Four students were assessed while teachers implemented DRC interventions focused on promoting sustained active engagement in academic tasks. In two cases (Students 5 and 7), sensitivity to change of the DBR-MIS Academic Engagement was positive and strong with *Tau-U* suggesting medium effects (i.e., *Tau-U* = 0.59 and 0.76, respectively) and Hedges' *g* resulting in very large values (i.e., *g* = 6.99 and 3.24, respectively). In the other two cases (Students 6 and 8), the magnitude of both estimates ranged

from trivial to small, and the direction of the effect for Student 6 was contrary to what was expected.

DBR-MIS Organizational Skills. The evaluation of sensitivity to change for the DBR-MIS Organizational Skills was based on eight students whose teachers considered them to be poorly organized and to have difficulties in following directions and arriving at class prepared to learn. Changes between baseline and intervention were documented for four cases (Students 1, 5, 12, and 13), meaning that *Tau-U* coefficients ranged from small to medium (i.e., *Tau-U* = 0.37, 0.64, 0.69, and 0.37, respectively), and Hedges' *g* was associated consistently with very large effects (i.e., *g* = 3.42, 3.11, 2.41, 2.80). The other four cases (Students 1, 4, 14, and 16) showed trivial changes in both estimates.

DBR-MIS Disruptive Behavior. Sensitivity to change of the DBR-MIS Disruptive Behavior was assessed by examining the changes in scores across seven students whose behaviors were potentially distracting to others or interfering with the learning of others. *Tau-U* and Hedges' *g* provided evidence for a significant reduction in scores within four cases (Students 2, 7, 8, and 21): *Tau-U* ranged from small to medium effect sizes (i.e., *Tau-U* = -0.22, -0.25, -0.38, and -0.52, respectively), and Hedges' *g* from medium to very large (i.e., *g* = -0.82, -0.56, -1.35, -1.65). Divergent results occurred for Student 18 for whom the magnitude of *Tau-U* was null, and the standardized mean difference indicated a small effect in the direction contrary to the expectations. Finally, two cases (Students 14 and 23) did not show relevant changes across the two estimates.

DBR-MIS Oppositional Behavior. Four students received DRC interventions designed to target patterns of behavior, including irritability, argumentativeness, defiance, and disrespect to adults. The DBR-MIS Oppositional Behavior was associated with significant effects on both

estimates for only one case (Student 16) in that *Tau-U* indicated a small effect (i.e., *Tau-U* = -0.25) and Hedges' *g* provided support for a medium change (i.e., *g* = -0.59). Two cases (Student 24 and 25) showed trivial changes. Finally, the direction of the effect was contrary to the expectations for Student 3 in that *Tau-U* indicated a positive but null effect, whereas Hedges' *g* was associated with a positive and medium effect (i.e., *g* = 0.68).

<Between-Case Effect Size Estimates>

Mixed effect meta-analysis. Four out of five DBR-MISs demonstrated sufficient sensitivity to change across participants. *Tau-U* coefficients were significant and consistent with the expected direction (see Table 6).

The highest effect sizes were associated with the DBR-MIS Engagement (*Tau-U* = 0.34, *SE* = 0.11, 95% CI = [0.12, 0.55], *p* <.001) and Organizational Skills (*Tau-U* = 0.28, *SE* = 0.07, 95% CI = [0.13, 0.42], *p* < .001). *Tau-U* was significant but smaller for the DBR-MIS Interpersonal Skills (*Tau-U* = 0.20, *SE* = 0.09, 95% CI = [0.01, 0.38], *p* = .04) and Disruptive Behavior (*Tau-U* = -0.21, *SE* = 0.07, 95% CI = [-0.35, -0.06], *p* = .01). Although the average effect calculated for DBR-MIS Oppositional Behavior was consistent with an expected reduction in these behaviors, mean *Tau-U* was trivial and not statistically significant (*Tau-U* = -0.04, SE = 0.09, 95% CI = [-0.21, 0.12], *p* = .60). In Table 6, random effects were not reported because heterogeneity indices were not significant, and the samples were relatively small. According to the literature, the fixed effect estimates are often preferable because they are less likely to be biased when conducted in studies with small numbers of observations (Greenland, 1994; Poole & Greenland, 1999; Schwartzer et al., 2015).

Two-level hierarchical linear model. We next used the combination of HLM and *BC-SMD* as an alternative approach to estimating sensitivity to change of the five DBR-MISs. Visual

analysis and within-case estimates suggested that outcome scores could be explained by the combination of fixed and random effects of cases, intervention, and linear trend (i.e., intervention by-time interaction). The reliability of the slope coefficients of the five scales was acceptable, varying between .60 and .72 (i.e., DBR-MIS Interpersonal Skills = .72, DBR-MIS Academic Engagement = .68, DBR-MIS Organizational Skills = .60, DBR-MIS Disruptive Behavior = .68, and DBR-MIS Oppositional Behavior = .69). Finally, the reader should note that the results of Model 4 were not included in the paper because the random effects associated with the linear trend were approximately 0 for all the scales.

DBR-MIS Interpersonal Skills. The results of the three models are reported in Table 7. When the fixed effect of the DRC intervention was included as the only predictor (Model 1), scores on the DBR-MIS Interpersonal Skills significantly improved by an average of 2.92 points ($t_{211} = 3.27$, $p = .001$). However, when the constraint of equal effect across cases was relaxed (Model 2), the fixed effect of intervention was no longer significant, and the presence of large random effects indicated non negligible variability across cases; in other words, differences in the outcome immediately after the intervention implementation varied a great deal as some students showed significant changes across the two phases, whereas others did not. Finally, the interaction between time and intervention (Model 3) was not significantly correlated to the outcome. Out of the three models, AIC and BIC associated with Model 2 had the smallest values, hence indicating the best model fit.

DBR-MIS Academic Engagement. Estimates of the three models are reported in Table 8. When considering the fixed effect of intervention alone (Model 1), scores on the DBR-MIS Engagement were associated with an immediate improvement by an average of 3.98 points ($t_{161} = 2.08$, $p = .04$). However, when considering random effects of the intervention (Model 2), the

fixed effect was no longer significant, and large variability was recorded across cases. Finally, a significant effect of linear trend within the intervention phase significantly correlated with the outcome (Model 3); on average, the DRC intervention contributed to improved scores on the DBR-MIS Academic Engagement by 1.94 points followed by an additional 0.24 points per day of intervention ($t_{160} = 4.28$, $p < .001$) for the DBR-MIS Academic Engagement. In other words, immediate behavioral changes following the intervention varied greatly between students; nonetheless, there were significant additional improvements on the scores of DBR-MIS Academic Engagement across cases. Out of the three models, AIC and BIC associated with Model 3 indicated the best model fit.

DBR-MIS Organizational Skills. The results of the three models are reported in Table 9. Similarly to the scale measuring academic engagement, the fixed effect of the intervention was associated with improved scores by an average of 3.75 points ($t_{363}= 2.75$, $p = .01$) when included in Model 1, but no longer significant when the assumption of equal intervention effect was relaxed (Model 2). Finally, when added to the set of predictors, a linear trend was significantly associated with the scores on the scale; on average, the DRC intervention contributed to improved scores on the DBR-MIS Organizational Skills by 2.57 points. This was followed by an additional 0.18 points per day of intervention ($t_{362} = 6.34$, $p < .001$). Random effects were still associated with the effect of intervention indicating large variability across cases. In other words, immediate behavioral changes following the intervention varied greatly between students, and significant additional improvements occurred across cases over time. Out of the three models, AIC and BIC associated with Model 3 indicated the best model fit.

DBR-MIS Disruptive Behavior. Table 10 reports the results of the three models for the scale. When estimated as a fixed effect, the DRC intervention contributed to significant

decreases in the scores on the DBR-MIS Disruptive Behavior obtained by the students (γ_1 = -2.61, t_{335} = -3.18, p = .002). However, when the constraint of equal effect across cases was relaxed, Model 2 no longer provided evidence of sensitivity to the effects of change of DBR-MIS Disruptive Behavior. Random effects associated with the intervention indicated large variability across the cases. When added to the model, the fixed effect of the trend was not significant (Model 3). AIC and BIC differences were inconsistent for DBR-MIS Disruptive Behavior with BIC supporting Model 1, whereas AIC supported Model 2.

DBR-MIS Oppositional Behavior. Table 11 reports the results of the three models for the scale. In Model 1, the fixed effect of the intervention was not significantly related to the outcome measured on DBR-MIS Oppositional Behavior. In addition, neither Model 2 nor Model 3 was associated with statistically significant coefficients. AIC and BIC were not associated with better model fit than the null model (i.e., where no predictors were considered).

BC-SMD estimates. Tables 7 through 11 include the *BC-SMD* estimates and standard errors calculated based on the coefficients and the total variance of the HLMs. The effect sizes were reported for all three models tested; however, we will comment only on the effect sizes associated with the best-fitting model for each DBR-MIS. Although results were consistent with the meta-analyses conducted on *Tau-U*, the overall magnitude of the effects was substantially larger.

A moderate effect size was found for the DBR-MIS Interpersonal Skills (*BC-SMD* = 0.76, SE = 0.46), very large effect sizes were found for DBR-MIS Academic Engagement (*BC-SMD* = 1.93, SE = 0.81) and Organizational Skills (*BC-SMD* = 1.80, *SE* = 0.51), and a small effect size was found for DBR-MIS Disruptive Behavior (*BC-SMD* = -0.40, *SE* = 0.15). By contrast, the effect size for the DBR-MIS Oppositional Behavior was null.

<Discussion>

The purpose of this study was to provide evidence for the sensitivity to change of five DBR-MISs designed to assess both academic enablers (i.e., interpersonal skills, academic engagement, organizational skills) and problem behaviors (i.e., disruptive and oppositional behaviors) in the context of a two-month DRC intervention. *Tau-U* and Hedges' *g* were calculated to quantify sensitivity to change of DBR-MIS within each AB single case study and used as relatively independent metrics to evaluate the presence of substantial intervention effects. Then, individual estimates were combined by using meta-analytic procedures. We found a high degree of correspondence between fixed effects of *Tau-U* and *BC-SMD* coefficients for all five scales: DBR-MIS Academic Engagement and Organizational Skills consistently demonstrated the strongest effects, followed by Disruptive Behavior and DBR-MIS Interpersonal Skills. The results were also consistent across metrics for DBR-MIS Oppositional Behavior in indicating that there was no effect captured by that scale. Finally, we found that the estimates for DBR-MIS Academic Engagement and Organizational Skills demonstrated a significant linear trend during intervention, supporting the hypothesis that the effects of the DRC intervention might occur gradually over time.

Results of this study make three main unique contributions to the development of progress monitoring tools in the context of DRC interventions. First, this study provides more robust evidence for the sensitivity to change of three DBR-MIS that were used in the pilot investigation (Hustus et al., 2020). We provided support for the use of DBR-MIS for monitoring student behavioral changes in response to DRC interventions targeting academic engagement and organizational skills. There was large variability associated with the changes immediately following the intervention implementation across cases. In addition, DBR-MIS Academic

Engagement and DBR-MIS Organizational Skills were sensitive to cumulative improvements of intervention over time. The presence of large variability associated with the intervention effect on DBR-MIS Academic Engagement indicates that the use of this scale might not capture immediate changes of the interventions, but be better suited to account for smaller changes over an extended period of time.

DBR-MIS Disruptive Behavior was also sensitive to the effects of DRC interventions in that substantial changes were found in the expected direction. Although large variability across students was associated with the effect of intervention, fit indices did not fully support the inclusion of random effects in the model.

Consistent with the examination of Hustus and colleagues (2020), the effects measured on DBR-MIS Oppositional Behavior were unconvincing both within and between cases. What remains unknown at this time, however, is whether the DBR-MIS Oppositional Behavior as a tool is insensitive to the effects of interventions or whether oppositional behaviors are less responsive to a DRC intervention.

A second major contribution of this paper involved the introduction of one new DBR-MIS for measuring interpersonal skills (Volpe et al., 2019). This scale has been designed to evaluate students' positive social behaviors while interacting with teachers or classmates. Results at the individual-level showed that the scale was sensitive to the effects of a two-month DRC intervention for two of the students involved in the study. Behavioral changes were substantial, followed the expected direction, and did not show a linear trend. However, when the variability between-cases was considered, the average effect of intervention no longer supported the sensitivity of the scale.

Although additional research for DBR-MIS Interpersonal Skills is necessary in order to corroborate evidence of its sensitivity to change, the assessment of interpersonal skills in the classroom setting is relevant because strong social competence leads to cooperative interactions with others, and is related to academic success and achievement test scores, even when controlling for potential confounding factors, such as school attendance, IQ, gender, ethnicity, and family composition (Gustavsen, 2017; Hall & DiPerna, 2017; Jenkins & Demaray, 2015; Lessard & Juvonen, 2018; Wentzel, 1993). By contrast, difficulties in interpersonal skills often represent goals for school-based behavioral interventions and are core features of a wide variety of disabilities, including ADHD, conduct disorder, autism spectrum disorders, and intellectual disabilities (Gresham, Sugai, & Horner, 2001). DiPerna, Volpe, and Elliott (2002) suggested that interpersonal skills along with engagement, study skills, and motivation, should be included in the broad category of academic enablers in that they play a crucial role in enhancing classroom learning and should be considered as part of a comprehensive assessment and intervention programs for students who are at risk for academic difficulties.

Third, sensitivity to change was calculated, accounting for the presence of time trends within the intervention phase. The inclusion of such a trend in the analyses was relevant because responses to behavioral interventions might occur gradually (Swan & Pustejovsky, 2018) or require consistency over time in order to be observed by raters (Whitcomb & Merrell, 2013). Although the five DBR-MIS were ranked similarly for sensitivity to change along with parametric and non-parametric estimates, *Tau-U* coefficients showed substantially lower magnitudes (i.e., from null to moderate) compared to *BC-SMD* (i.e., from null to very large). Such a difference is likely related to how the two metrics are calculated and interpreted. First, the inclusion of the intervention phase in calculating the *Tau-U* coefficient introduces additional

variance that likely reduces the results (Parker et al., 2011). By contrast, effect sizes that include intervention-by-time as a predictor within the intervention phase (e.g., *BC-SMD*) are equal or greater than the coefficients estimated for models wherein intervention is the only predictor (Parker et al., 2011); specifically, the magnitude of the coefficient increases when data points in the intervention phase are linearly or quadratically distributed. Moreover, the distribution for an SMD effect size with single-case data is not presently known. Thus, users have been cautioned about the use of Cohen's interpretive rules of thumb as these may drastically overestimate intervention effects. For example, Pustejovsky (2019) suggested that meaningful effects are likely higher based on outcomes reported in the intervention literature.

Over the last decade, scholars have controlled for the intervention trend in a very limited number of single case studies (e.g., Klingbeil et al., 2019; Parker et al., 2011). The investigation conducted by Klingbeil and colleagues (2019) represents one of the few studies wherein the intervention trend was considered in educational settings. The authors examined the effects of curriculum-based measures of reading (CBM-R) on reading fluency in a sample of 88 AB comparisons across 29 studies. *Tau-U* coefficients were calculated correcting both for baseline and intervention phase trend (i.e., $Tau\text{-}U_{\text{full}}$); this estimate was selected because interventions targeting fluency are likely to take time before producing changes, hence improvements might be observed later on in the intervention phase. Therefore, $Tau\text{-}U_{\text{full}}$ represents a relevant estimate for measuring academic outcomes in that professionals often aim to evaluate the changes both in level and in the trend of student performance effects produced by target interventions. Similar considerations might apply to interventions for social-emotional outcomes, such as academic enablers and externalizing symptoms.

In a recent review on the *Tau-U* family of coefficients, Brossart and colleagues (2018) suggested that the decision to control for the trend, whether within the intervention or within the baseline phase, should be based on both theoretical and empirical rationale. On the one hand, theoretical rationale involves considerations about experimental design and previous evidence from the literature in the field. For instance, controlling for the intervention trend is often appropriate in settings wherein interventions might produce gradual changes on the target outcome (Valentine et al., 2016); by contrast, controlling for baseline trend might be appropriate for the evaluation of intervention effectiveness in addition to co-occurrent interventions or when reactivity effects (e.g., Hawthorne effect; Brown, 1992) are expected. On the other hand, the trend correction may rely on an empirical rationale. For instance, a slope coefficient equal to or greater than .40 has been suggested as an indicator of a significant trend (Parker et al., 2011). However, when relying on a few number of observations, the slope coefficients might reach statistical levels of significance because of outliers or random distribution of the data.

Finally, the nature of target behaviors and treatment integrity might help to explain differences observed in sensitivity to change across the five DBR-MIS. We suggest that DRC interventions might differently affect the magnitude and the trend of behavioral outcomes. Teachers might promote adaptive behaviors more successfully than they prevent students from exhibiting negative behaviors towards adults, peers, and materials; also, the lack of sensitivity to change of DBR-MIS Oppositional Behavior was not totally unexpected in that interventions targeting these types of behaviors are more likely to be effective when the family is actively involved (Markward & Bride, 2001). Such interventions might require that adults interrupt the cycle of coercive interactions that have become automated in children with oppositional behaviors (and breaking this cycle may be more effective when there is consistency from home

and school). Also, the coercive interpersonal cycle is such that teachers may feel punished when they attempt to change these behaviors (e.g., behaviors get worse or children are more negatively reactive) (Carr, Taylor, & Robinson, 1991). Moreover, considerable variability across single studies and the lack of consideration for the duration of the intervention leave open the possibility of different results in future examinations.

Furthermore, sensitivity to change might depend on the degree to which the intervention is implemented with treatment integrity. Intuitively, interventions can be linked to student behavioral changes only if teacher behaviors are consistent with its principles and strategies. Data from the TIF demonstrated that interventions were carried out with a sufficient level of adherence. Owens and colleagues (2020) hypothesized that the minimum required level of treatment integrity for a DRC intervention might vary based on the severity of student behavioral problems; specifically, 51% integrity might be enough to produce observable outcomes in students who are not diagnosed with clinical disorder, whereas higher levels of adherence might be necessary for students with severe psychopathologies. Interestingly, the one DBR-MIS that did not demonstrate sensitivity to change (i.e., DBR-MIS Oppositional Behavior) was also associated with lower levels of DRC treatment integrity (i.e., 51%) than the other four scales. This combination of more intense target behaviors and lower levels of DRC treatment integrity may have meant that the intervention was less effective for addressing these types of behaviors.

<Limitations and Direction for Future Research>

Although the results provided evidence for the use of four DBR-MIS in the context of progress monitoring, it is worth noting that the study presented a few limitations. First, the evaluation of sensitivity to change was limited to two variables, in that only treatment integrity and changes on DBR-MIS scores were considered. However, several other variables (e.g.,

malleability, severity of behavioral disorders) might play a role in determining whether a DBR-MIS can capture the effects of behavioral interventions. The extent to which such variables can be operationalized varies; for example, the inclusion of a malleability index relative to a behavioral domain in response to a specific intervention might be relatively easy to obtain (e.g., expert consensus), whereas less obvious might be to provide with a consensual agreement on how to establish the severity of behavioral disorders across students. We did use a promising screening tool (i.e., ITRF) in order to identify students with behavioral needs, but its scores are not necessarily indicators of severity on a specific domain. Future studies are needed in order to support sensitivity to change of DBR-MIS when controlling for additional variables.

Second, although we selected four of the most advanced metrics available to calculate sensitivity to change, they are not exempt from limitations; in fact, the usage of any effect statistic is inherently biased towards certain characteristics of the effects of behavioral interventions. For instance, after being proposed by Parker and colleagues (2011), *Tau-U*$_{\text{A vs. B + trend B}}$ has never been used in a research study; therefore, no strong evidence has been provided regarding its psychometric features, such as comprehensive examination of its distribution or the presence of ceiling effects. Our findings demonstrated that such a metric never reached 1, though the upper bound of confidence intervals was beyond that value in four cases. This result is obviously impossible for a measure bounded between -1 and 1, hence further psychometric investigations are needed. Another limitation was related to the effects estimated in order to calculate *BC-SMD*. Although HLM allows for the estimation of both fixed and random effects, we limited the focus of our analyses to the fixed effects within the intervention phase due to small sample sizes for each scale; however, the inclusion of random effects for one or more predictors might be one of the most promising directions for future studies.

Third, we limited the investigation of sensitivity to change to the degree to which a measure is capable of detecting changes in behavior in response to the implementation of an evidence-based intervention (Husted, Cook, Farewell, & Gladman, 2000). This approach has been referred to as a single-measure, internal responsiveness (Husted et al., 2000) or distribution-based (Stratford et al., 1996) approach. However, a second approach – known as comparative, external responsiveness (Husted et al., 2000), or criterion-based (Stratford et al., 1996) – has been suggested for the investigation of sensitivity to change. Within this approach, it is most common to compare changes detected in an experimental method to those observed in an established reference measure. Future studies are needed in order to identify whether behavioral changes on the DBR-MIS correspond to those detected by existing assessment methods.

Fourth, we calculated sensitivity to change of DBR-MIS by considering both level and slope and assuming the presence of similar effects on the outcome; however, the findings might point in a different direction in that some scales showed significant shifts in the average level while others significant linear trends. The absence of apriori hypotheses regarding the expected effects is common in the evaluation of intervention responsiveness because investigators tend to be more interested in describing the presence of effects rather than how the intervention affects the outcome over time. Moreover, the results demonstrated that four of the DBR-MIS are sensitive to the effects of DRC interventions. We caution against generalizing these findings when other treatments are used (e.g., pharmacological; Daniels et al., 2017). Further research is needed in order to establish sensitivity to change and its patterns for different behaviors and when alternative interventions are put into place.

Fifth, important methodological limitations of the current study include the use of a series of AB designs and of small sample sizes when each DBR-MIS is considered separately.

Evidence for sensitivity to change was drawn based on the rating behavior of four to eight teachers, depending on the scale. In addition, in the absence of experimental design, regression to the mean might affect the data used in the analyses. However, at least three factors would constitute negative evidence for this interpretation of the results: (a) the effectiveness of DRC interventions relies on a strong literature base (e.g., Fabiano et al., 2017), (b) student behaviors were rated on multiple occasions prior to the implementation of the intervention (see Table 1), and (c) on average, variability increased across the two phases (see Table 4). However, future studies employing experimental procedures (e.g., ABAB, reversal design, or Randomized Controlled Trial) or collecting data from larger sample sizes will further strengthen the evidence for sensitivity to change of the DBR-MIS presented in this paper.

<Conclusions>

The present study was the first to provide evidence for sensitivity to change of three DBR-MIS in the context of progress monitoring response to a common classroom intervention. These DBR-MIS have been designed to allow professionals working with early elementary students to monitor interventions targeting academic engagement, organizational skills, and disruptive behaviors. Sensitivity to change was partially supported for the newly introduced DBR-MIS Interpersonal Skills; however, more evidence is needed due to large differences across cases. The results did not support sensitivity to change of the DBR-MIS Oppositional Behavior, though we were not able to establish whether such a tool was not capable to capture the effects of the intervention or oppositional behaviors were not responsive to a DRC intervention.

Sensitivity to change of DBR-MIS was measured in the context of a two-month intervention conducted on a daily basis. The length and the frequency of the intervention made this study particularly relevant in that the pattern was examined in relation to changes in outcome

level and trend. DBR-MIS Academic Engagement and Organizational Skills exhibited gradual improvements over time, whereas changes captured through the DBR-MIS Disruptive Behavior were constant over the course of the classroom intervention. Combined with evidence from empirical patterns of DRC interventions, these findings enable professionals to examine whether the intervention is effective and to anticipate reasonable expectations from future intervention sessions.

Finally, we encourage test developers and professionals to consider the examination of sensitivity to change of progress monitoring tools in their practice. On the one hand, when discussing the psychometric adequacy of formative assessment measures, empirical evidence should be provided regarding its likelihood of detecting a clinically substantial treatment effect (Guyatt et al., 1987). On the other hand, when implementing an intervention, only tools that are sensitive to the effects of change allow for the collection of information that teachers and psychologists can use for data-driven decisions in applied settings.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Baek, E. K., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2016). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation*, *24*, 590–606. https://doi.org/10.1080/096202011.2013.835740.

Benner, G. J., Kutash, K., Nelson, J. R., & Fisher, M. B. (2013). Closing the achievement gap of youth with emotional and behavioral disorders through multi-tiered systems of support. *Education and Treatment of Children*, *36*(3), 15–29. https://doi.org/10.1353/etc.2013.0018

Borenstein, M., Higgins, J., Hedges, L., & Rothstein, H. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Briesch, A. M., Riley-Tillman, T. C., & Chafouleas, S. M. (2016). *Direct Behavior Rating: Linking assessment, communication, and intervention*. New York, NY: Guilford Press.

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau-U for single-case experimental designs. *Cogent Psychology*, *5*(1). https://doi.org/10.1080/23311908.2018.1518687

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, *2*, 141–178. http://dx.doi.org/10.1207/s15327809jls0202_2

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. http://dx.doi.org/10.1177/0049124104268644

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Byiers, B., Reichle, J., & Symons, F. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology*, *21*, 397–414. http://dx.doi.org/10.1044/1058-0360(2012/11-0036

Carr, E. G., Taylor, J. C., & Robinson, S. (1991). The effects of severe behavior problems in children on the teaching behavior of adults. *Journal of Applied Behavior Analysis*, *24*, 523–535. http://dx.doi.org/ 10.1901/jaba.1991.24-523

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data: What they tell us and what they don't. *Behavior modification*, *37*, 378–390. http://dx.doi.org/10.1177/0145445513476609

Chafouleas, S. M., Kilgus, S. P., Jaffery, R., Riley-Tillman, T. C., Welsh, M., & Christ, T. J. (2013). Direct behavior rating as a school-based behavior screener for elementary and middle grades. *Journal of School Psychology*, *51*, 367–385. http://dx.doi.org/10.1016/j.jsp.2013.04.002

Chafouleas, S. M., Riley-Tillman, T. C., & Christ T. J. (2009) Direct behavior rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention*, *34*, 195–200. http://dx.doi.org/10.1177/1534508409340391

Chafouleas, S. M., Riley-Tillman, T. C., & Sugai, G. (2007). *School-based behavioral assessment: Informing instruction and intervention*. New York, NY: Guilford Press.

Chafouleas, S. M., Sanetti, L. M., Kilgus, S. P., & Maggin, D. M. (2012). Evaluating sensitivity to behavioral change using direct behavior rating single-item scales. *Exceptional Children*, *78*, 491–505. http://dx.doi.org/10.1177/001440291207800406

Chafouleas, S. M., Volpe, R. J., Gresham, F. M., & Cook, C. R. (2010). School-based behavioral assessment within problem-solving models: Current status and future directions. *School Psychology Review*, *39*, 343–349.

Christ, T. J. (2006). Short term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of slope to construct confidence intervals. *School Psychology Review*, *35*, 128–133.

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. http://dx.doi.org/10.1177/1534508409340390

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, C. R., Volpe, R. J., & Delport, J. (2013). A review of systematic monitoring in EBD: The promise of change sensitive brief behavior rating scales. In H. W. Walker & F. M. Gresham (Eds.). *Handbook of evidence-based practices for emotional and behavioral disorders: Applications in schools* (pp. 211–228). New York, NY: Guilford.

Daniels, B., Briesch, A. M., Volpe, R. J., & Owens, J. S. (2019). Content Validation of Direct Behavior Rating Multi-Item Scales for Assessing Problem Behaviors. *Journal of*

Emotional and Behavioral Disorders, 1063426619882345. https://doi.org/10.1177/1063426619882345

Daniels, B., Volpe, R. J., Briesch, A. M., & Fabiano, G. A. (2014). Development of a problem-focused behavioral screener linked to evidence-based intervention. *School Psychology Quarterly*, *29*(4), 438–451. http://dx.doi.org/10.1037/spq0000100

Daniels, B., Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2017). Dependability and treatment sensitivity of Multi-Item Direct Behavior Rating Scales for interpersonal peer conflict. *Assessment for Effective Intervention*, *43*, 48–59. http://dx.doi.org/10.1177/1534508417698456

Dart, E. H., Arora, P. G., Collins, T. A., & Doll, B. (2019). Progress monitoring measures for internalizing symptoms: A systematic review of the peer-reviewed literature. *School Mental Health*, *11*, 265–275. http://dx.doi.org/10.1007/s12310-018-9299-7

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2002). A model of academic enablers and elementary reading/language arts achievement. *School Psychology Review*, *31*, 298–312.

Elliott, S. N., Malecki, C. K., & Demaray, M. K. (2001). New directions in social skills assessment and intervention for elementary and middle school students. *Exceptionality, 9*, 19–32. https://dx.doi.org/10.1207/S15327035EX091&2_3

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Fabiano, G. A., Pyle, K., Kelty, M. B., & Parham, B. R. (2017). Progress monitoring using direct behavior rating single item scales in a multiple-baseline design study of the daily report card intervention. *Assessment for Effective Intervention*, *43*, 21–33. http://dx.doi.org/10.1177/1534508417703024

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. http://dx.doi.org/10.1007/s11121-007-0070-9

Greenland, S. (1994). Invited commentary: a critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, *140*(3), 290–296. https://dx.doi.org/10.1093/oxfordjournals.aje.a117248

Gresham, F. M. (2005). Response to intervention: An alternative means of identifying students as emotionally disturbed. *Education and Treatment of Children*, *28*, 328–344.

Gresham, F. M., Sugai, G., & Horner, R. H. (2001). Interpreting outcomes of social skills training for students with high-incidence disabilities. *Exceptional Children*, *67*, 331–344. http://dx.doi.org/10.1177/001440290106700303

Gustavsen, A. M. (2017). Longitudinal relationship between social skills and academic achievement in a gender perspective. *Cogent Education*, *4*(1), 1–16. http://dx.doi.org/10.1080/2331186x.2017.1411035

Guyatt, G., Walter, S., & Norman, G. (1987). Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases*, *40*, 171–178. http://dx.doi.org/10.1016/0021-9681(87)90069-5

Hall, G. E., & DiPerna, J. C. (2017). Childhood social skills as predictors of middle school academic adjustment. *The Journal of Early Adolescence*, *37*, 825–851. http://dx.doi.org/10.1177/0272431615624566

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Higgins, J.P.T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley Online Library.

Hintze, J. M., & Marcotte, A. M. (2010). Student assessment and data-based decision making. In T. A. Glover & S. Vaughn (Eds.), *The promise of response to intervention: Evaluating current science and practice* (pp. 57–77). New York, NY: Guilford Press.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47.

Husted, J. A., Cook, R. J., Farewell, V. T., & Gladman, D. D. (2000). Methods for assessing responsiveness: a critical review and recommendations. *Journal of Clinical Epidemiology*, *53*, 459–468.

Hustus, C. L., Owens, J. S., Volpe, R. J., Briesch, A. M., & Daniels, B. (2020). Treatment Sensitivity of Direct Behavior Rating–Multi-Item Scales in the Context of a Daily Report Card Intervention. *Journal of Emotional and Behavioral Disorders*, *28*(1), 29–42. https://dx.doi.org/10.1177/1063426618806281

Jenkins, L. N., & Demaray, M. K. (2015). An investigation of relations among academic enablers and reading outcomes. *Psychology in the Schools*, *52*, 379–389. http://dx.doi.org/10.1002/pits.21830

Jensen-Doss, A., Haimes, E. M. B., Smith, A. M., Lyon, A. R., Lewis, C. C., Stanick, C. F., & Hawley, K. M. (2018). Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*, 48–61. http://dx.doi.org/10.1007/s10488-01607630

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Welsh, M. E. (2012). Direct Behavior Rating scales as screeners: A preliminary investigation of diagnostic accuracy in

elementary school. *School Psychology Quarterly*, *27*, 41–50. http://dx.doi.org/10.1037/a0027150

Kilgus, S. P., Riley-Tillman, T. C., & Kratochwill, T. R. (2016). Establishing interventions via a theory-driven single case design research cycle. *School Psychology Review*, *45*, 477–498. http://dx.doi.org/10.17105/SPR45-4.477-498

Kilgus, S. P., Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., & Welsh, M. E. (2014). Direct Behavior Rating as a school-based behavior universal screener: Replication across sites. *Journal of School Psychology*, *52*, 63–82. http://dx.doi.org/10.1016/j.jsp.2013.11.002

Klingbeil, D. A., Van Norman, E. R., McLendon, K. E., Ross, S. G., & Begeny, J. C. (2019). Evaluating Tau-U with oral reading fluency data and the impact of measurement error. *Behavior Modification*, *43*, 413–438. http://dx.doi.org/10.1177/0145445518760174

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124–144. http://dx.doi.org/10.1037/14376-003

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*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Lessard, L. M., & Juvonen, J. (2018). Losing and gaining friends: Does friendship instability compromise academic functioning in middle school?. *Journal of School Psychology*, *69*, 143–153. http://dx.doi.org/10.1016/j.jsp.2018.05.003

Markward, M. J., & Bride, B. (2001). Oppositional defiant disorder and the need for family-centered practice in schools. *Children & Schools*, *23*, 73–83. http://dx.doi.org/10.1093/cs/23.2.73

Miller, F. G., Crovello, N. J., & Chafouleas, S. M. (2017). Progress monitoring the effects of daily report cards across elementary and secondary settings using Direct Behavior Rating: Single Item Scales. *Assessment for Effective Intervention*, *43*, 34–47. http://dx.doi.org/10.1177/1534508417691019

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A monte carlo simulation study. *The Journal of Experimental Education*, *84*, 473–509. http://dx.doi.org/10.1080/00220973.2015.1065216

Owens, J. S., Evans, S. W., Coles, E. K., Holdaway, A. S., Himawan, L. K., Mixon, C. S., & Egan, T. E. (2020). Consultation for classroom management and targeted interventions: examining benchmarks for teacher practices that produce desired change in student behavior. *Journal of Emotional and Behavioral Disorders*, *28*(1), 52–64. http://dx.doi.org/10.1177/1063426618795440

Owens, J. S., Holdaway, A. S., Zoromski, A. K., Evans, S. W., & Himawan, L. K., Girio-Herrera, E., & Murphy, C. E. (2012). Incremental benefits of a daily report card intervention over time for youth with disruptive behavior. *Behavior Therapy*, *43*, 848–861. http://dx.doi.org/ 10.1016/j.beth.2012.02.002

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284–299. http://dx.doi.org/10.1016/j.beth.2010.08.006

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D,, & R Core Team. (2019). *nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-140*, https://CRAN.R-project.org/package=nlme.

Poole, C., & Greenland, S. (1999). Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology*, *150*(5), 469–475. https://dx.doi.org/10.1093/oxfordjournals.aje.a010035

Prasad, V., Brogan, E., Mulvaney, C., Grainge, M., Stanton, W., & Sayal, K. (2013). How effective are drug treatments for children with ADHD at improving on-task behaviour and academic achievement in the school classroom? A systematic review and meta-analysis. *European Child & Adolescent Psychiatry*, *22*, 203–216. http://dx.doi.org/10.1007/s00787-012-0346-x

Pustejovsky, J. E. (2016). *scdhlm: Estimating hierarchical linear models for single-case designs. R package version 0.3*, http://github.com/jepusto/scdhlm.

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99–112. http://dx.doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*, 217–235. http://dx.doi.org/10.1037/met0000179

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kaufmann, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of Special Education* (2nd ed.). New York, NY: Routledge.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-Comparable Effect Sizes in Multiple Baseline Designs. *Journal of Educational and Behavioral Statistics*, *39*(5), 368–393. http://dx.doi.org/10.3102/1076998614547577.

Pustejovsky, J. E., & Swan, D. M. (2017). *SinglecaseES: A calculator for single-case effect size indices, R package version 0.3*. Retrieved from https://github. com/jepusto/SingleCaseES.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164. http://dx.doi.org/10.2307/271063

Riley-Tillman, T. C., & Burns, M. K. (2009). *The Guilford practical intervention in the schools series. Evaluating educational interventions: Single-case design for measuring response to intervention*. New York, NY: Guilford Press.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software*, *45*(4), 1–20. http://dx.doi.org/10.18637/jss.v045.i04

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*, 79–88.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. http://dx.doi.org/10.1214/aos/1176344136

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. New York, NY: Springer.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research (NCER 2015-002)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*, 385–405. http://dx.doi.org/10.1037/a0032964

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Stratford, P. W., Binkley, J., Solomon, P., Finch, E., Gill, C., & Moreland, J. (1996). Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Physical Therapy*, *76*, 359–365. http://dx.doi.org/10.1093/ptj/76.4.359

Swan, D. M., & Pustejovsky, J. E. (2018). A gradual effects model for single-case designs. *Multivariate Behavioral Research*, *53*, 574–593. http://dx.doi.org/10.1080/00273171.2018.1466681

Tarlow, K. R. (2017, March). *Tau-U for single-case research (R code)*. Retrieved from http://ktarlow.com/stats/

Valentine, J. C., Tanner-Smith, E. E., Pustejovsky, J. E., & Lau, T. S. (2016). *Between-case standardized mean difference effect sizes for single-case designs: A primer*. Oslo, Norway: The Campbell Collaboration.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*, 142–151. http://dx.doi.org/10.1080/17489530802505362

Vannest, K. J., Davis, J. L., Davis, C. R., Mason, B. A., & Burke, M. D. (2010). Effective intervention for behavior with a daily behavior report card: A meta-analysis. *School Psychology Review*, *39*, 654–672.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item direct behavior rating scales for engagement and disruptive behavior. *School Psychology Review*, *41*, 246–261.

Volpe, R. J., & Briesch, A. M. (2015). Multi-item direct behavior ratings: Dependability of two levels of assessment specificity. *School Psychology Quarterly*, *30*, 431-442. http://dx.doi.org/10.1037/spq0000115

Volpe, R. J., & Briesch, A. M. (2016). Dependability of two scaling approaches to direct behavior rating multi-item scales assessing disruptive classroom behavior. *School Psychology Review*, *45*, 39–52. http://dx.doi.org/10.17105/SPR45-1.39-52

Volpe, R. J., Chaffee, R. K., Yeung, T. S., & Briesch, A. M. (2020). Initial Development of Multi-item Direct Behavior Rating Measures of Academic Enablers. *School Mental Health*, *12*(1), 77–87. https://dx.doi.org/10.1007/s12310-019-09338-w

Volpe, R. J., & Fabiano, G. A. (2013). *Daily behavior report cards: An evidence-based system of assessment and intervention*. New York, NY: Guilford Press.

Volpe, R. J., & Gadow, K. D. (2010). Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. *School Psychology Review*, *39*, 350–363

Volpe, R. J., Gadow, K. D., Blom-Hoffman, J., & Feinberg, A. B. (2009). Factor-analytic and individualized approaches to constructing brief measures of ADHD behaviors. *Journal of Emotional and Behavioral Disorders*, *17*, 118–128. http://dx.doi.org/10.1177/1063426608323370

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the Direct Observation Form. *School Psychology Review*, *38*, 382–401.

von der Embse, N. P., Scott, E. C., Kilgus, S. P. (2015). Sensitivity to change and concurrent validity of direct behavior ratings for academic anxiety. *School Psychology Quarterly*, *30*, 244–259. http://dx.doi.org/10.1037/spq0000083

Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology*, *85*, 357–364. http://dx.doi.org/10.1037/0022-0663.85.2.357

Whitcomb, S. A., & Merrell, K. W. (2013). *Behavioral, social, and emotional assessment of children and adolescents* (4th ed.). New York, NY: Routledge.

Wickham H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Table 1

Descriptive Statistics of the Duration[a] *of DRC Intervention by Behavioral Domain Weighted by Cases*

		Baseline					Intervention				
Scale	*N*[b]	*M*	*SD*	min	max	CR[c]	*M*	*SD*	min	max	CR[c]
Interpersonal Skills	4	14.00	6.48	7	20	87.50	40.00	2.16	38	43	85.63
Academic Engagement	4	10.00	7.35	6	21	82.50	31.50	11.09	22	43	85.71
Organizational Skills	8	11.25	6.84	6	24	88.89	35.25	8.35	21	43	80.14
Disruptive Behavior	7	14.43	8.68	7	31	89.11	34.57	9.98	22	53	75.21
Oppositional Behavior	4	16.25	6.85	8	24	84.61	39.00	2.94	35	42	85.26

[a] Duration refers to the number of school days
[b] Number of students evaluated on DBR-MIS
[c] Completion Rate (%)

Table 2

Treatment Integrity Observations of DRC Implementation Procedures Weighted by Cases

	Observation Time (minutes)		Treatment Integrity[a]		Ratings[b]		
DRC goals	*M*	*SD*	*M*	*SD*	*M*	min	max
Interpersonal Skills	42.72	9.55	74.04	6.52	8.25	6	12
Academic Engagement	53.26	10.99	74.90	30.43	5.40	3	7
Organizational Skills	36.87	7.44	67.17	26.38	6.75	3	12
Disruptive Behavior	51.35	15.25	69.06	21.89	5.67	3	8
Oppositional Behavior	38.16	9.31	53.50	26.29	7.60	6	12

[a] Scores range from 53.50 to 74.90.

[b] Scores range from 5.40 to 8.25.

Table 3

DBR-MIS Descriptive Statistics by Phase Weighted by Cases before and After Multiple Imputations

	Before Multiple Imputations						After Multiple Imputations					
	Baseline			Intervention			Baseline			Intervention		
Scale	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Interpersonal Skills[a]	12.51	2.71	11.97	16.00	3.51	14.16	12.25	2.31	12.32	15.93	3.23	14.88
Academic Engagement[a]	12.18	3.33	14.03	18.78	5.33	18.12	12.44	3.09	14.01	18.42	5.07	17.55
Organizational Skills[a]	14.49	3.69	14.79	20.42	5.97	21.00	14.35	3.56	14.67	19.77	5.83	19.92
Disruptive Behavior[b]	16.92	4.43	15.05	13.32	6.49	11.32	16.76	4.21	14.68	13.97	6.10	13.07
Oppositional Behavior[b]	12.00	5.64	7.79	11.89	8.69	7.70	12.74	5.85	8.84	11.61	7.77	8.20

[a]Higher scores indicate higher demonstration of these skills.

[b]Higher scores indicate greater severity of these problems.

Table 4

Changes of DBR-MIS Ranges Across Phases Weighted by Cases

	Baseline		Intervention		Differences	
Scale	min	Max	min	max	min	max
Interpersonal Skills	5.33	18.90	6.64	26.90	1.31	8.00
Academic Engagement	8.58	14.90	9.42	27.60	0.84	12.70
Organizational Skills	8.40	20.00	9.74	26.80	1.34	6.80
Disruptive Behavior	7.40	24.90	7.18	20.40	-0.22	-4.50
Oppositional Behavior	3.38	23.60	4.76	23.40	1.38	-0.20

Table 5

Within-Case Non Overlap Tau-U and Standardized Mean Difference

		Tau-U				Hedges' *g*			
Scale	Student	Est	*SE*	CI_{low}	CI_{upp}	Est	*SE*	CI_{low}	CI_{upp}
Interpersonal Skills									
	1	0.16	0.24	-0.32	0.63	2.55	0.74	1.11	4.00
	2	0.41	0.21	0.00	0.82	1.55	0.48	0.61	2.49
	3	-0.11	0.17	-0.44	0.21	-0.51	0.32	-1.14	0.12
	4	0.36	0.16	0.04	0.68	0.70	0.29	0.13	1.28
Academic Engagement									
	5	0.59	0.25	0.09	1.08	6.99	1.83	3.39	10.58
	6	-0.12	0.29	-0.68	0.44	-0.11	0.48	-1.06	0.83
	7	0.76	0.25	0.27	1.26	3.24	0.88	1.51	4.96
	8	0.20	0.16	-0.12	0.52	0.45	0.55	-0.63	1.52
Organizational Skills									
	5	0.64	0.25	0.15	1.13	3.42	0.96	1.55	5.30
	6	0.00	0.28	-0.56	0.55	0.05	0.48	-0.89	0.99
	1	0.37	0.24	-0.10	0.85	3.11	0.84	1.47	4.75
	12	0.69	0.22	0.26	1.13	2.41	0.64	1.16	3.65
	13	0.37	0.22	-0.06	0.80	2.80	0.70	1.44	4.16
	14	0.11	0.24	-0.35	0.58	0.06	0.44	-0.80	0.92
	4	0.16	0.16	-0.16	0.48	0.09	0.26	-0.42	0.59
	16	0.12	0.15	-0.18	0.42	-0.09	0.33	-0.75	0.57
Disruptive Behavior									
	7	-0.25	0.25	-0.74	0.25	-0.56	0.41	-1.36	0.24
	18	-0.03	0.21	-0.45	0.39	0.46	0.36	-0.25	1.16
	14	0.07	0.23	-0.39	0.52	-0.16	0.35	-0.85	0.53
	2	-0.22	0.21	-0.64	0.19	-0.82	0.34	-1.48	-0.15
	21	-0.52	0.20	-0.92	-0.12	-1.65	0.61	-2.84	-0.46
	8	-0.38	0.16	-0.69	-0.06	-1.35	0.35	-2.03	-0.67
	23	-0.03	0.17	-0.37	0.30	0.04	0.23	-0.41	0.50
Oppositional Behavior									
	24	0.03	0.23	-0.42	0.49	-0.10	0.33	-0.75	0.55
	25	-0.08	0.18	-0.44	0.27	-0.21	0.26	-0.73	0.31
	3	0.18	0.16	-0.14	0.50	0.68	0.34	0.02	1.34

16	-0.25	0.15	-0.55	0.05	-0.59	0.27	-1.12	-0.06

Table 6

Fixed Effects and Hetereogeneity Estimates for DBR-MIS Resulting from Meta-analysis Across Students

Scale	Est	*SE*	CI_{low}	CI_{upp}	*z*	*p*	tau^2	*H*	I^2	*Q*	*df*	*p*
Interpersonal Skills	0.20	0.09	0.01	0.38	2.09	.04	0.03	1.34	0.44	5.35	3	0.15
Academic Engagement	0.34	0.11	0.12	0.55	3.07	<.001	0.07	1.54	0.58	7.10	3	0.07
Organizational Skills	0.28	0.07	0.13	0.42	3.79	<.001	0.01	1.14	0.23	9.05	7	0.25
Disruptive Behavior	-0.21	0.07	-0.35	-0.06	-2.78	.01	0.01	1.07	0.13	6.90	6	0.33
Oppositional Behavior	-0.04	0.09	-0.21	0.12	-0.52	.60	0.01	1.15	0.25	3.99	3	0.26

Table 7

DBR-MIS Interpersonal Skills Model Parameters and Goodness of Fit

	Model 0	Model 1		Model 2		Model 3	
Parameter	Estimate	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects							
Intercept (γ_0)		13.22***	1.47	12.47***	1.15	12.46***	1.15
Intervention (γ_1)		2.92***	0.89	3.65†	2.12	2.57	2.21
Time x Intervention (γ_2)						0.05	0.03
Random effects							
Intercept (η_0)		2.48	0.00	1.85	0.03	1.85	0.00
Intervention (η_1)				3.94	0.05	3.90	0.01
Residuals		4.60	0.00	4.32	0.01	4.32	0.00
Effect size							
Adjusted (*BC-SMD*)		0.55	0.18	0.76	0.46	0.89	0.47
degrees of freedom		211		211		210	
B-A						31	
Model fit indexes							
Akaike Info. Criterion	1286.13	1276.21		1266.05°		1271.51	
Bayesian Info. Criterion	1299.61	1293.04		1289.61°		1298.40	

Note. Model 0 = null model (i.e., random intercept only); Model 1 = fixed and varying intercepts and fixed treatment effect; Model 2 = fixed and varying intercepts and intervention; Model 3 = fixed and varying intercepts and fixed treatment effect. Model 4 was not reported in the paper because fixed and random coefficients did not differ from Model 3.

†$p \leq .10$. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

° Smallest fit index across the four models.

Table 8

DBR-MIS Academic Engagement Model Parameters and Goodness of Fit

	Model 0	Model 1		Model 2		Model 3	
Parameter	Estimate	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects							
Intercept (γ_0)		13.90***	1.92	13.01***	1.58	11.86***	1.53
Intervention (γ_1)		3.98*	1.92	4.64	3.08	1.94	3.37
Time x Intervention (γ_2)						0.24***	0.06
Random effects							
Intercept (η_0)		0.12	0.08	1.27	0.18	2.17	0.02
Intervention (η_1)				5.05	0.38	6.09	0.02
Residuals		6.44	0.01	5.05	0.05	4.32	0.00
Effect size							
Adjusted (*BC-SMD*)		0.60	0.31	0.87	0.63	1.93	0.81
degrees of freedom		161		161		160	
B-A						32	
Model fit indexes							
Akaike Info. Criterion	956.93	951.32		950.85		937.57°	
Bayesian Info. Criterion	969.36	966.82		972.55		962.32°	

Note. Model 0 = null model (i.e., random intercept only); Model 1 = fixed and varying intercepts and fixed treatment effect; Model 2 = fixed and varying intercepts and intervention; Model 3 = fixed and varying intercepts and fixed treatment effect. Model 4 was not reported in the paper because fixed and random coefficients did not differ from Model 3.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

° Smallest fit index across the four models.

Table 9

DBR-MIS Organizational Skills Model Parameters and Goodness of Fit

	Model 0	Model 1		Model 2		Model 3	
Parameter	Estimate	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects							
Intercept (γ_0)		15.96***	1.41	14.42***	1.07	13.92***	1.11
Intervention (γ_1)		3.75**	1.36	5.27**	2.13	2.57	2.20
Time x Intervention (γ_2)						0.18***	0.03
Random effects							
Intercept (η_0)		0.56	0.40	2.28	0.04	2.68	0.00
Intervention (η_1)				5.58	0.04	5.83	0.00
Residuals		6.52	0.04	4.40	0.01	3.98	0.00
Effect size							
Adjusted (*BC-SMD*)		0.55	0.21	1.05	0.44	1.80	0.51
degrees of freedom		363		363		362	
B-A						35	
Model fit indexes							
Akaike Info. Criterion	2193.48	2189.60		2145.58		2116.22°	
Bayesian Info. Criterion	2209.14	2209.16		2172.97		2147.51°	

Note. Model 0 = null model (i.e., random intercept only); Model 1 = fixed and varying intercepts and fixed treatment effect; Model 2 = fixed and varying intercepts and intervention; Model 3 = fixed and varying intercepts and fixed treatment effect. Model 4 was not reported in the paper because fixed and random coefficients did not differ from Model 3.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

° Smallest fit index across the four models.

Table 10

DBR-MIS Disruptive Behavior Model Parameters and Goodness of Fit

	Model 0	Model 1		Model 2		Model 3	
Parameter	Estimate	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects							
Intercept (γ_0)		16.87***	1.80	16.82***	1.60	16.88***	1.60
Intervention (γ_1)		-2.61**	0.82	-2.62†	1.43	-1.70	1.57
Time x Intervention (γ_2)						-0.05	0.04
Random effects							
Intercept (η_0)		4.38	0.00	3.84	0.05	3.85	0.01
Intervention (η_1)				3.17	0.07	3.21	0.02
Residuals		4.56	0.00	4.34	0.03	4.32	0.00
Effect size							
Adjusted (*BC-SMD*)		-0.40	0.15	-0.44	0.26	-0.66	0.32
degrees of freedom		335		335		334	
B-A						42	
Model fit indexes							
Akaike Info. Criterion	1994.37	1979.40		1972.66°		1978.76	
Bayesian Info. Criterion	2009.71	1998.55°		1999.48		2009.40	

Note. Model 0 = null model (i.e., random intercept only); Model 1 = fixed and varying intercepts and fixed treatment effect; Model 2 = fixed and varying intercepts and intervention; Model 3 = fixed and varying intercepts and fixed treatment effect. Model 4 was not reported in the paper because fixed and random coefficients did not differ from Model 3.

†$p \leq .10$. *$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

° Smallest fit index across the four models.

Table 11

DBR-MIS Oppositional Behavior Model Parameters and Goodness of Fit

	Model 0	Model 1		Model 2		Model 3	
Parameter	Estimate	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects							
Intercept (γ_0)		13.26***	3.28	13.39***	2.43	13.38***	2.42
Intervention (γ_1)		-0.26	1.11	-0.34	1.69	-0.10	1.98
Time x Intervention (γ_2)						-0.02	0.05
Random effects							
Intercept (η_0)		6.28	0.00	4.52	0.07	4.50	0.02
Intervention (η_1)				2.61	0.08	2.66	0.02
Residuals		5.82	0.00	5.72	0.02	5.72	0.00
Effect size							
Adjusted (*BC-SMD*)		-0.03	0.13	-0.04	0.24	-0.09	0.26
degrees of freedom		216		216		215	
B-A						35	
Model fit indexes							
Akaike Info. Criterion	1409.56°	1409.46		1410.01		1414.60	
Bayesian Info. Criterion	1423.14°	1426.41		1433.74		1441.68	

Note. Model 0 = null model (i.e., random intercept only); Model 1 = fixed and varying intercepts and fixed treatment effect; Model 2 = fixed and varying intercepts and intervention; Model 3 = fixed and varying intercepts and fixed treatment effect. Model 4 was not reported in the paper because fixed and random coefficients did not differ from Model 3.

$^{\dagger}p \leq .10$. $^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

° Smallest fit index across the four models.

Appendix A

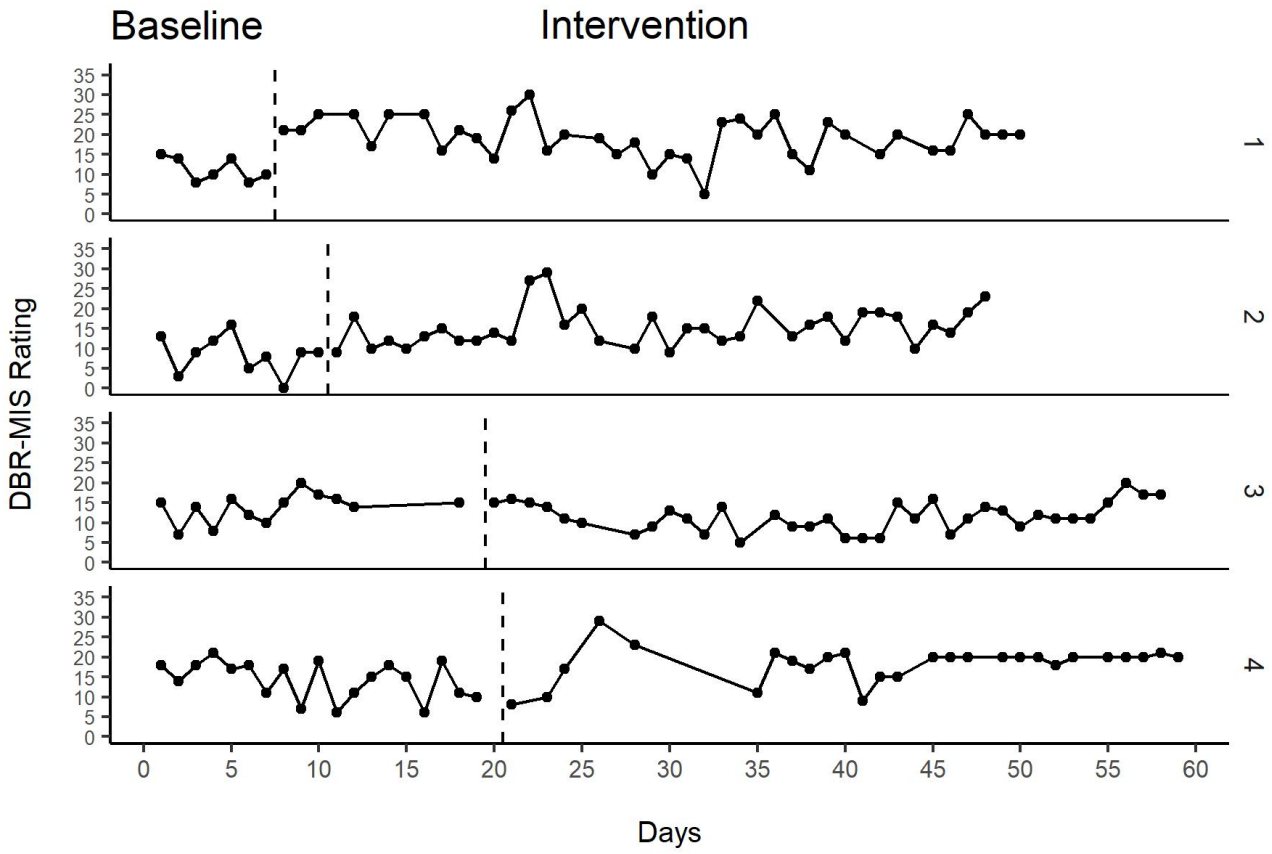

Figure 1. Interpersonal Skills

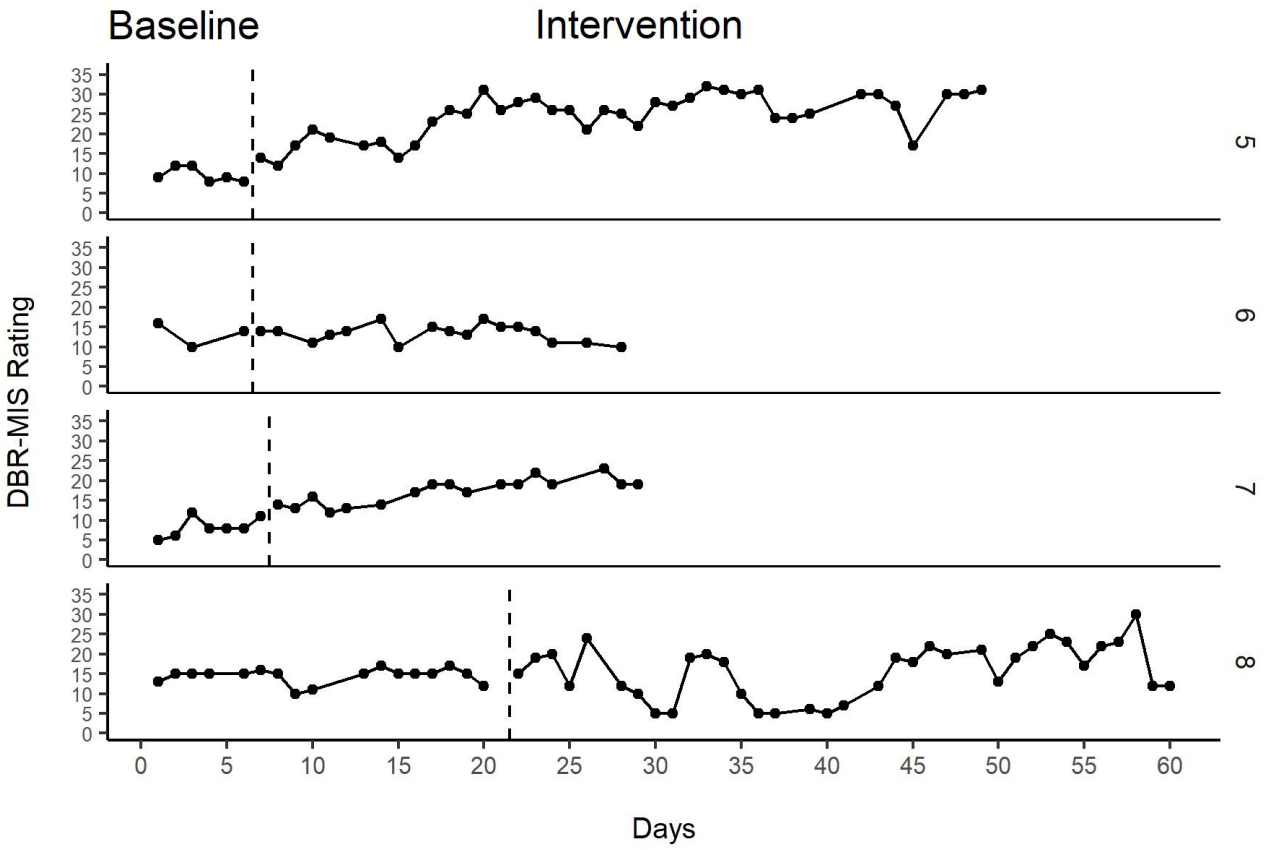

Figure 2. Academic Engagement

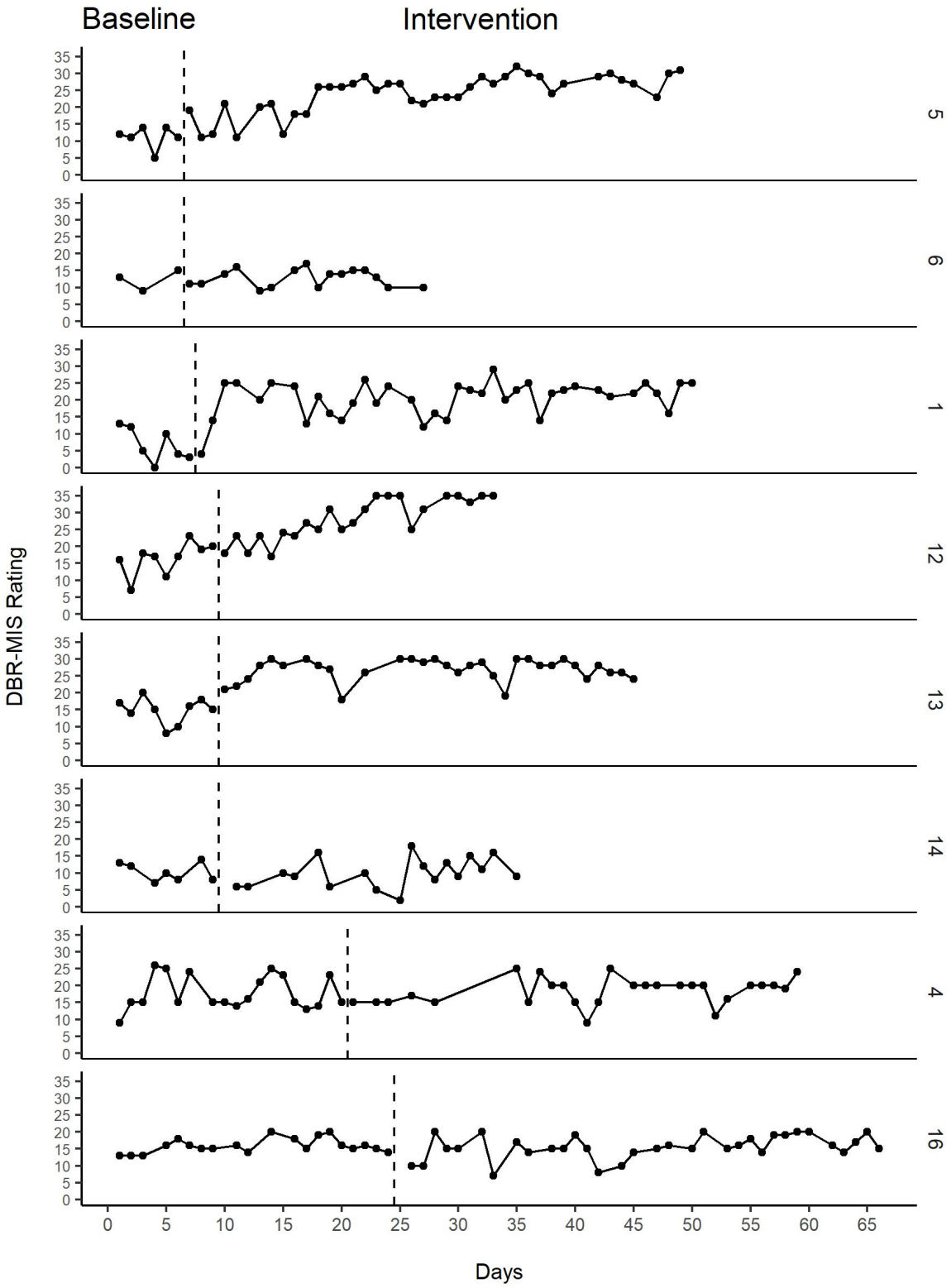

Figure 3. Organizational Skills

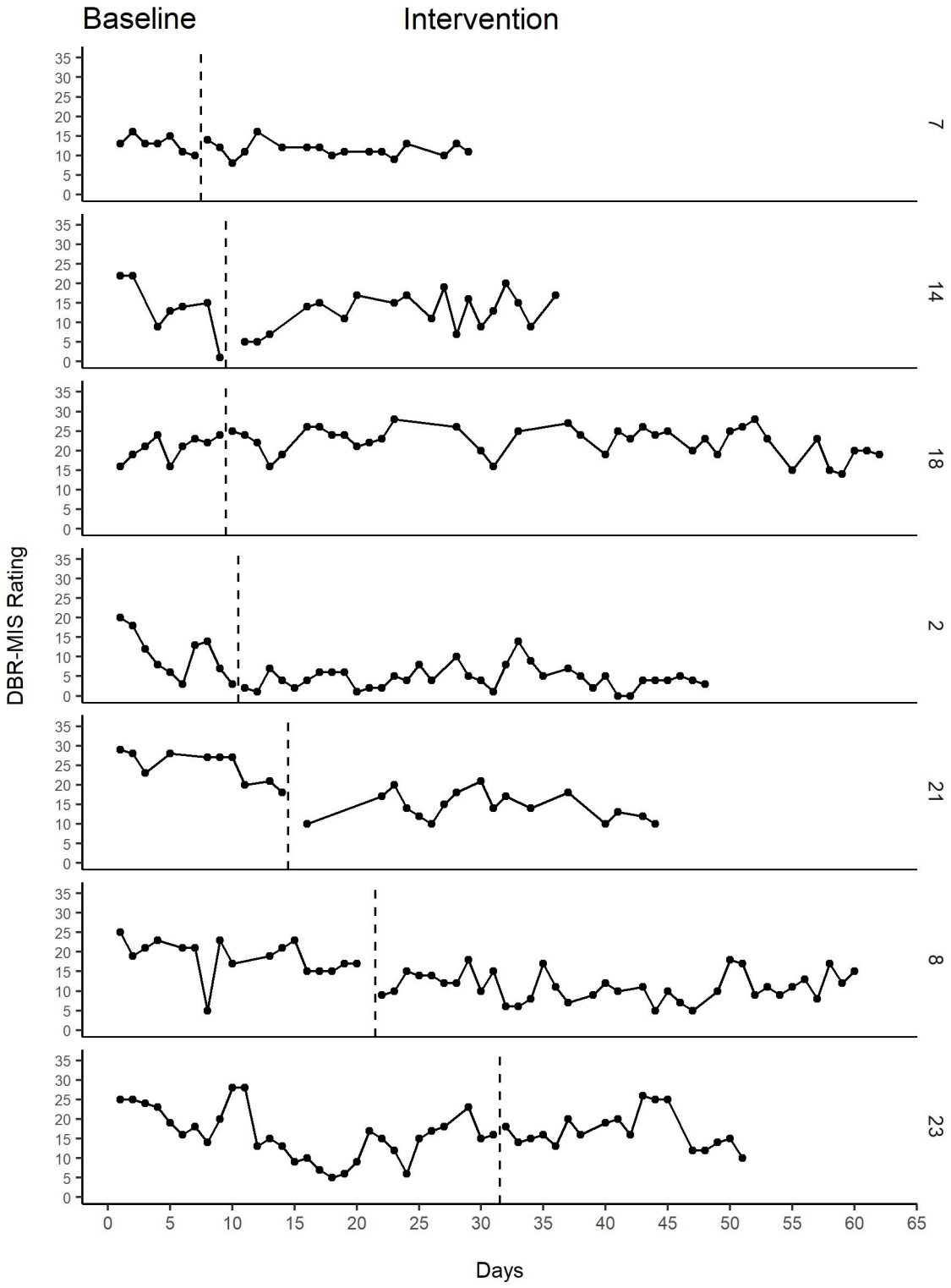

Figure 4. Disruptive Behavior

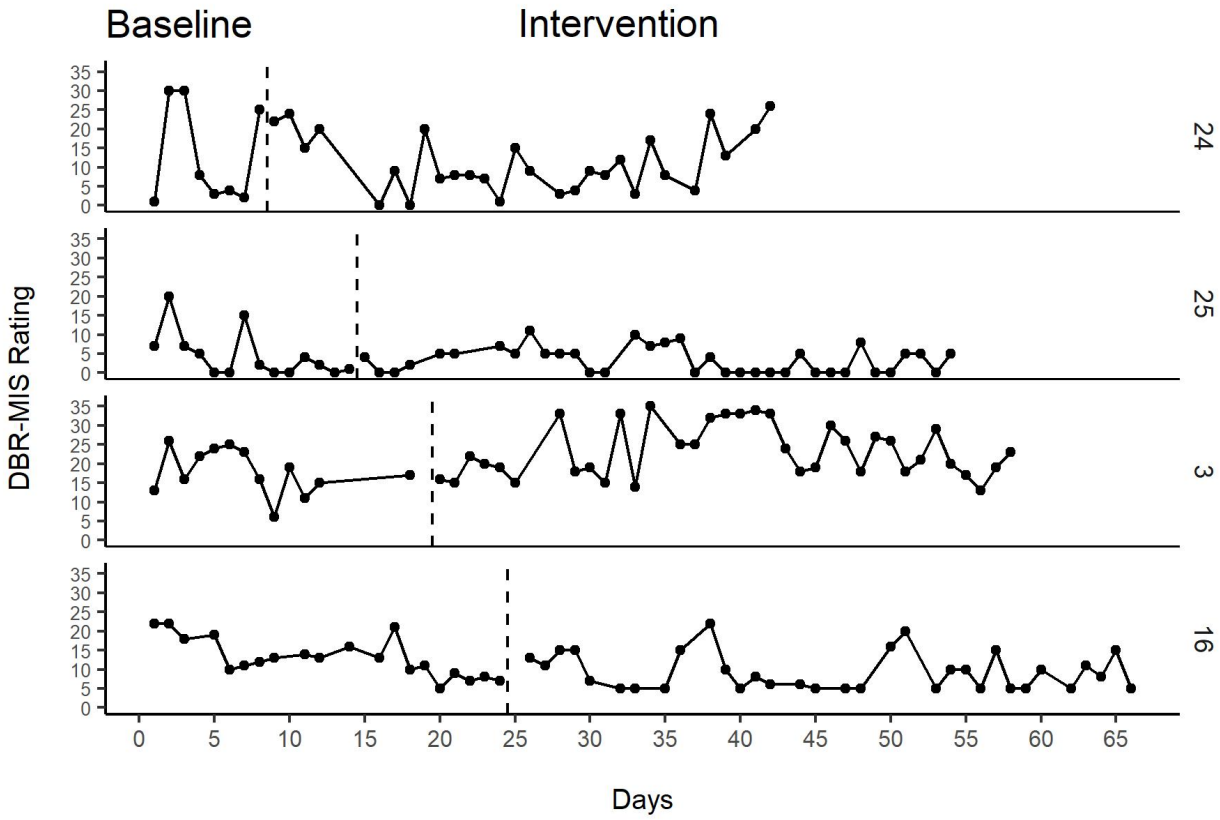

Figure 5. Oppositional Behavior