Hammill Institute on Disabilities

Assessment for Effective Intervention
2017, Vol. 42(2) 119–126

DOI: 10.1177/1534508416658007
aei.sagepub.com
SAGE

Direct Behavior Rating Instrumentation: Evaluating the Impact of Scale Formats

Faith G. Miller, PhD[1], T. Chris Riley-Tillman, PhD[2], Sandra M. Chafouleas, PhD[3], and Alyssa A. Schardt, MA[1]

Abstract

The purpose of this study was to investigate the impact of two different Direct Behavior Rating–Single Item Scale (DBR-SIS) formats on rating accuracy. A total of 119 undergraduate students participated in one of two study conditions, each utilizing a different DBR-SIS scale format: one that included percentage of time anchors on the DBR-SIS scale and an explicit reference to duration of the target behavior (percent group) and one that did not include percentage anchors nor a reference to duration of the target behavior (no percent group). Participants viewed nine brief video clips and rated student behavior using one of the two DBR-SIS formats. Rating accuracy was determined by calculating the absolute difference between participant ratings and two criterion measures: systematic direct observation scores and DBR-SIS expert ratings. Statistically significant differences between groups were found on only two occasions, pertaining to ratings of academically engaged behavior. Limitations and directions for future research are discussed.

Keywords

behavior, assessment, direct behavior ratings, accuracy

More than 35 years ago, Saal, Downey, and Lahey (1980) published their influential manuscript in *Psychological Bulletin* titled "Rating the Ratings: Assessing the Psychometric Quality of Rating Data," which reviewed measurement challenges in obtaining reliable and valid ratings of performance and behavior. Although the primary focus of their manuscript involved issues related to response biases (e.g., halo effect, leniency, or severity), the authors heeded that there was much work to be done "to maximize the desirable psychometric characteristics of ratings and minimize or eliminate the undesirable characteristics" (p. 426). Consequently, a wealth of research has been undertaken regarding measurement issues associated with ratings in behavioral research; numerous potential sources of measurement error have been identified (for a comprehensive summary, see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), and among those acknowledged are those associated with characteristics of the scales. More specifically, the very format of the scales and scale anchors can introduce systematic measurement error into behavioral ratings (Tourangeau, Rips, & Rasinski, 2000). Indeed, the issue of scaling has long been of interest to psychometricians (e.g., Likert, 1932; Stevens, 1946; Thurstone, 1928). Consequently, the purpose of the present investigation was to examine the instrumentation of Direct Behavior Rating–Single Item Scales (DBR-SIS) by comparing rating accuracy associated with two different scale formats.

Scale Development of DBR-SIS

DBR-SIS were developed as a systematic approach to monitor the progress of student behavior. DBR-SIS evolved from a myriad of user-created DBR tools (behavior report cards, homeschool notes, point sheets), where "home grown" assessments were created to fill the void of available behavioral progress monitoring tools. Yet, the psychometric properties of these tools remain unknown. To fill this gap, DBR-SIS were designed as a hybrid behavioral assessment method, combining the benefits of systematic direct observation (SDO) with an efficient rating scale format (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Chafouleas, Riley-Tillman, & Sugai, 2007). Specifically, teachers observe students during a pre-specified observation period and immediately following the observation, record ratings of the target behaviors on a 0 to 10 scale. As outlined by Chafouleas (2011), considerable work was undertaken to inform the development of the

[1]University of Minnesota, Minneapolis, USA
[2]University of Missouri, Columbia, USA
[3]University of Connecticut, Storrs, USA

Corresponding Author:
Faith G. Miller, Department of Educational Psychology, University of Minnesota, 250 Education Sciences Building, Minneapolis, MN 55455, USA.
Email: fgmiller@umn.edu

scales, including reviews of prior research on scale development and validation studies. In regard to scale development, various options were considered in terms of polarity (unipolar or bipolar), the scale of measurement (level of measurement or range of scores), the scale format (continuous, pseudocontinuous, or categorical), and number of gradients included (see Christ & Boice, 2009). Prior research examining variations in scale format (e.g., number of gradients, continuous or categorical scales) supported the flexibility of the scales and suggested negligible differences in rating accuracy as a function of scale format (Briesch, Kilgus, Chafouleas, Riley-Tillman, & Christ, 2012; Christ, Riley-Tillman, & Chafouleas, 2009; Riley-Tillman, Christ, Chafouleas, Boice-Mallach, & Briesch, 2011). However, questions remain regarding the impact of other key features of the scales that have yet to be studied empirically.

Although DBR-SIS demonstrates flexibility in some aspects, it is not clear whether other features of the scale affect measurement error. In particular, duration was selected as the behavioral dimension of interest in developing DBR-SIS, and the scale includes explicit instructions referencing duration (i.e., place a mark along the line that best reflects the percentage of total time the student exhibited each target behavior). Furthermore, percentage of time anchors is included on the scale (0%, 50%, and 100%). Several bodies of literature, briefly described below, informed this selection including (a) measurement research related to scale construction and (b) behavioral observation research (e.g., Hartley, Trueman, & Rodgers, 1984; Newstead & Arnold, 1989; Skinner, Rhymer, & McDaniel, 2000).

Although the structure of DBR-SIS resembles a Likert-type scale, the inclusion of percentage anchors results in equal interval measurement not typically associated with Likert-type scales. For example, a 4-point Likert-type scale might include the following response options: never, sometimes, often, always, and are often considered to be at the ordinal level of measurement. Conversely, DBR-SIS includes a numerical range from 0 to 10 that can be considered to be at the interval level of measurement. The result is a pseudocontinuous scale, as ratings are rounded to the nearest whole number for interpretation. This presumably results in more desirable measurement characteristics, given increased precision of equal interval measurement compared with the ordinal measurement associated with traditional Likert-type scales (Hartley et al., 1984). Prior measurement research on the issue of scaling has supported the use of percentage anchors when possible, as opposed to numerical anchors or verbal anchors (such as those typically utilized on Likert-type scales) as they are often less accurate (Hartley et al., 1984; Newstead & Arnold, 1989). In addition, it has been argued that including percentage of time rather than general frequency descriptors allows for a more accurate comparison of results between individuals (Diener, Smith, & Fujita, 1995).

Although global recommendations have been made regarding scale construction, an additional contextual consideration in the measurement of behavior involves the target behaviors of interest. The process of selecting and operationalizing behavior targets and selecting measurement approaches is inherently related (Skinner et al., 2000). That is, in behavioral observational research, operationalizing the target behavior informs the dimension of interest and subsequent measurement of that behavior. Efforts have been made to assist in the classification of behaviors and selecting the appropriate measurement approach.

Altmann (1974) and Saudargas and Lentz (1986) conceptualized behavior as falling along a state-event continuum. Within this framework, state behaviors are considered to be behaviors with meaningful duration (e.g., on-task, out of seat, etc.) whereas event behaviors are of brief duration and have clear start and end points (e.g., calling out, swearing, etc.). Many behaviors can be conceptualized as either states or events, but some fall within both categories. In fact, it can be argued that some of the core behavioral competences targeted on DBR-SIS forms do not clearly fall into one category—while academically engaged (AE) behavior can be readily categorized as a state, disruptive (DB) and respectful (RS) behavior could be conceptualized as either states or events. The categorization of state versus event behaviors becomes increasingly important when deciding how to measure different behaviors. According to Saudargas and Lentz (1986), state behaviors are best studied by examining the percent of total time the behavior occurred across an observation session. Event behaviors, on the contrary, are best studied by examining their frequency. Because some of the behaviors on the DBR-SIS scale are not clearly interpreted as either states or events, it remains unclear how the scale format might affect rating accuracy.

Purpose

In light of these considerations, additional research is needed to evaluate the instrumentation of DBR-SIS, and the scaling in particular. Consequently, this study was designed to evaluate whether or not including percentage of time anchors and an explicit reference to duration on a DBR-SIS 0 to 10 scale affects the accuracy with which individuals rate student behavior. The following research question guided our inquiry:

> **Research Question 1:** Does the inclusion of percentage of time anchors and an explicit reference to duration on the DBR-SIS scale affect rating accuracy?

It was hypothesized that including percentage of time anchors would affect ratings only when the target behavior

was a clear state behavior (AE) but not for behaviors that did not clearly fall into one category (i.e., DB and RS).

Method

Participants

Participants registered online for one of four 1-hr study sessions, thus blindly self-selecting into one of two conditions that utilized different DBR-SIS scale formats: one who completed DBR-SIS with percentage of time anchors and were explicitly directed to rate duration (percent group) and one who completed DBR-SIS ratings without percentage of time anchors and were not explicitly directed to rate duration (no percent group). At the start of the session, basic demographic information was collected about the sample. The total sample was comprised of 119 undergraduate student participants enrolled in a large public university located in the southeastern United States. In sum, 55 students blindly self-selected into the percent group, whereas 64 self-selected into the no percent group. Ethnicity of the pool of potential subjects was identified as 64% White, 24% Black, 6% Asian, and 6% Other. Sixty-one percent of the subject pool were female. The majority of participants were freshman (71%) and aged 17 to 29 years. By participating in the study, participants partially fulfilled a research participation requirement for an introductory psychology course. All study procedures were completed in compliance with university human subjects review board policies.

Materials

Video clips. Ten researcher-created video clips of a simulated elementary classroom were developed for use in research studies, and were filmed in a first-grade classroom. To create the clips, a first-grade teacher was asked to engage in activities typical for an elementary setting, and eight elementary-age students (four male, four female) were asked to display typical prosocial classroom behavior, unless instructed otherwise. Prior to filming, target students were instructed to engage in various levels of AE, DB, and RS behavior. For this study, two target students were selected, one male and one female, and were positioned to provide an unobstructed view of their behavior. Because rating accuracy could vary as a function of the behavior exhibited in the clip, each clip was systematically constructed to display a student engaged in either a low, medium, or high level of each target behavior. Target students were coached by the research staff to act disruptively, disrespectfully, or disengaged for predefined periods of time. Graduate students coded the videos using the Multiple Option Observation System for Experimental Studies (MOOSES; Tapp, 2002), which incorporates a real-time event coding procedure using a 1-s coding interval. The MOOSES data were then used to categorize the clips into low, medium, or high levels; the behavior was deemed to occur at a low level if displayed during 0% to 25% of the clip, medium if displayed during 26% to 74% of the clip, or high if displayed ≥75% of the clip. Each clip was 60 s in duration, and was projected on a large screen for participants to view simultaneously. One of the clips was used during an initial training exercise, and the other nine were used as test stimuli. In five of the clips, participants were asked to rate the male student, whereas in four of the clips they were asked to rate the female student. All participants rated all nine clips: low AE, medium AE, high AE, low DB, medium DB, high DB, and low RS, medium RS, and high RS. The nine video clips were randomly ordered into a pre-set sequence that subsequently remained the same across study conditions.

DBR-SIS. Rating packets were developed for each condition that included DBR-SIS forms for each video clip. The standard DBR-SIS rating form was used for both conditions, which is comprised of three single-item scales, one for each target behavior of AE, DB, and RS behavior. Each scale is comprised of a single 100 mm line with 11 equal gradients (0–10) and qualitative anchors of "never," "sometimes," and "always" displayed at the beginning, middle, and end of the scale, respectively. Definitions and examples of each target behavior were provided on the top of the DBR-SIS form. Academically engaged behavior was defined as actively or passively participating in the classroom activity. Disruptive behavior was defined as student action that interrupts regular school or classroom activity. RS behavior was defined as compliant and polite behavior in response to adult direction and/or interactions with peers and adults. Beyond these standard elements of the rating form, two aspects were systematically altered depending on the condition the participant was assigned to. One group completed DBR-SIS ratings with percentage of time anchors (0%, 50%, 100%) on the scale (percent group), whereas the percentage of time anchors were removed from the scale for the other group (no percent group). The directions on the forms also differed by group. For the percent group, directions at the top of the DBR-SIS form were as follows: Place a slash (/) on the line through the number that best reflects the *percentage of total time* the target child exhibited the specified behavior during the observation session. Conversely, the following directions were printed on the forms for the no percent group: Place a slash (/) on the line through the number that best reflects *the extent to which* the target child exhibited the specified behavior during the observation session. This resulted in two different scaling approaches: (a) one that explicitly referenced duration and included percentage of time anchors of 0%, 50%, and 100%, and (b) a more global frequency scale with the percentage of time anchors removed that is consistent with those utilized in psychometric research (e.g., Rohrmann,

2007). Rating packets were created that contained nine 0 to 10 DBR-SIS scales, one rating form for each video clip.

Procedures

Study sessions spanned 1 hr. At the beginning of the session, training presentations were delivered by a graduate research assistant to (a) teach participants about rating student behavior using DBR-SIS, (b) describe procedures for rating behavior, (c) describe operational definitions and examples of the three target behavioral outcomes evaluated in this study (i.e., AE, RS, and DB), and (d) observe a sample video clip to demonstrate decisions made when rating student behavior using DBR-SIS. The percent group's presentation contained examples of DBR scales with percentage anchors and provided explanations that used duration of time as a frame of reference when rating. The no percent group's presentation did not use percentage indicators on the DBR scales or use duration of time as a frame of reference. Instead, they were instructed to rate the extent to which the student exhibited the target behavior, and explanations provided during the training included references to "slight," "moderate," or "high" levels of behavior displayed. After the training, participants watched the first video clip and immediately provided ratings of behavior on the DBR-SIS form in their rating packet. This process continued until all nine clips were rated and the study session concluded.

Data Analysis

To evaluate the extent to which including percentage of time anchors influenced rating accuracy, a series of analyses were conducted comparing DBR-SIS ratings to two external criteria to determine the degree of rater error associated with each DBR scale format (percent or no percent). As noted by Chafouleas, Kilgus, Riley-Tillman, Jaffery, and Harrison (2012), selecting an objective criterion against which to evaluate the accuracy of DBR-SIS is challenging; therefore, we utilized the method used in that study. Specifically, rater error scores were calculated by subtracting participants' DBR-SIS scores from scores obtained using two criteria: (a) SDO ($|DBR_{Part} - SDO|$) and (b) expert ratings ($|DBR_{Part} - DBR_{Exp}|$). Absolute scores were used in all analyses. Because no gold standard exists in relation to determining rater accuracy using DBR-SIS, a dual-pronged approach was deemed most appropriate for evaluating differences in rating accuracy between groups. Expert DBR-SIS scores were obtained using a consensus building procedure with 13 researchers with expertise in behavioral assessment (see Jaffery et al., 2015 for a full description of procedures), and utilizing the median expert DBR-SIS rating for comparison. Conversely, SDO scores were obtained by rounding and transforming the data obtained using MOOSES to the DBR-SIS scale. For example, behavior that was deemed to occur 73% of the time using MOOSES would be transformed to a rating of 7 on DBR-SIS.

Prior to analysis, all rater error scores were checked for accuracy of data entry and examined with regard to parametric assumptions. No cases were missing data, thus all cases were included in the analysis. Normality was evaluated through a review of skewness and kurtosis statistics, histograms, and formal tests of normality. Based on review of these data, the assumption of normality was not deemed tenable; skewness and kurtosis statistics were divided by their respective standard error estimates and all values except for Medium AE, DB, and RS behavior fell outside of |2.5|, and all tests for normality (Kolmogorov–Smirnov and Shapiro–Wilk) were statistically significant. Given the nature of these data and the presence of substantially skewed and leptokurtic distributions, nonparametric analyses were deemed most appropriate for evaluating differences between groups.

Results

In Table 1, the descriptive statistics for the DBR-SIS ratings for all nine behaviors by group (percent and no percent) are displayed. To examine differences in rater error scores, a series of Mann–Whitney *U* tests were performed to evaluate differences in mean ranks between groups. Holm–Bonferroni corrections were used to adjust for multiple comparisons across the nine clips. Results from the analyses revealed statistically significant differences between groups for both low AE and high AE clips. For the low AE clip, significant differences were found between groups using both SDO error scores ($|DBR_{Part} - SDO|$; $U = 1{,}185.5$, $p = .002$) and expert error scores ($|DBR_{Part} - DBR_{Exp}|$; $U = 889.5$, $p < .001$). Similarly, significant differences were found between groups on the high AE clip using both SDO error scores ($U = 1279.0$, $p = .004$) and expert error scores ($U = 1279$, $p = .004$). No statistically significant differences were found for the other clips.

To explore the practical significance of the significant findings, effect sizes were generated using the probability of superiority (PS) metric recommended by Nussbaum (2014). This metric is used to determine the probability that a randomly sampled score from one group would be higher than the comparison group. According to interpretive guidelines developed by Grissom (1994) and Grissom and Kim (2005), a PS value of 56% represents a small effect, 65% represents a medium effect, and 71% represents a large effect. For the low AE clip, PS values ranged from 66% ($|DBR_{Part} - SDO|$) to 75% ($|DBR_{Part} - DBR_{Exp}|$), suggesting a medium to large effect; a 66% to 75% chance of randomly drawing a higher error score for the no percent group on that clip. For the high AE clip, PS values were 64% for both $|DBR_{Part} - SDO|$ and $|DBR_{Part} - DBR_{Exp}|$ error scores, indicating a small to medium effect; a 64% chance

Table 1. Descriptive Statistics.

Behavior	Level	SDO rating	Expert rating	Group	*n*	*Mdn* DBR-SIS	IQR	
							Q1	Q3
Academically engaged	Low	1.7	1.0	Percent	64	1.00	0.25	2.00
				No percent	55	3.00	2.00	5.00
	Medium	4.8	4.0	Percent	64	6.00	5.00	8.00
				No percent	55	6.00	4.00	8.00
	High	9.8	10.0	Percent	64	10.00	9.00	10.00
				No percent	55	9.00	8.00	10.00
Disruptive	Low	1.7	2.0	Percent	64	3.00	1.00	5.00
				No percent	55	2.00	1.00	5.00
	Medium	5.7	7.0	Percent	64	8.00	6.00	9.00
				No percent	55	8.00	6.00	9.00
	High	8.5	9.0	Percent	64	9.00	8.25	10.00
				No percent	55	9.00	8.00	10.00
Respectful	Low	1.7	1.0	Percent	64	1.00	1.00	2.00
				No percent	55	2.00	1.00	3.00
	Medium	5.3	3.0	Percent	64	2.00	1.00	4.00
				No percent	55	3.00	2.00	5.00
	High	10.0	10.0	Percent	64	9.00	7.00	9.00
				No percent	55	9.00	7.00	10.00

Note. SDO = systematic direct observation; DBR-SIS = Direct Behavior Rating–Single Item Scale; IQR = interquartile range.

of randomly drawing a higher error score for the no percent group on that clip.

To derive robust estimates of confidence intervals for the sample, bootstrapping procedures were used. To this end, participants' scores are used to generate bootstrap samples, and a median is calculated for each sample. The bootstrapped medians are then put in order, from lowest to highest, and the central 95% of values are used to form the confidence interval (Keselman, Algina, Lix, Wilcox, & Deering, 2008). Bootstrapping procedures were used on the sample of rater error scores for all nine clips using IBM SPSS statistical computing software version 22. Ten thousand samples were created with replacement to generate 95% bootstrapped confidence intervals about the median error scores for each group (see Tables 2 and 3).

Discussion

Overall, findings from the present investigation suggest small but potentially important differences in rating accuracy between groups. Specifically, Low and High AE were the only two behavior clips that produced statistically significant differences in rating accuracy across the percent and no percent groups. In both cases, the percent group produced more accurate ratings than the no percent group. This finding was consistent across both SDO error scores and expert error scores, providing important corroboration in the absence of a clear gold standard by which to measure the accuracy of DBR-SIS ratings.

Considering previous research on state versus event behaviors, it is reasonable to expect that AE behavior would be the most affected by the scale format on the rating scale due to its clear classification as a state behavior (Saudargas & Lentz, 1986). That is, state behaviors are best measured by duration. Furthermore, academic engagement does not lend itself well to other possible dimensions of interest, such as frequency or intensity. That is, duration is unequivocally the most important dimension to measure, thus the scale referencing duration was associated with more accurate scores. Conversely, significant differences were not observed between scale formats for DB and RS behavior, both of which do not clearly fall into either category in the state-event dichotomy. As such, including the percent anchors and referencing duration did not appear to affect the accuracy of ratings, despite research pointing to inaccuracies associated with low-specificity descriptors such as those used in the no percent group (Lucas, Diener, & Larsen, 2009). Although we can only speculate as to the reasons for these discrepancies, it seems reasonable to expect that scale format may have greater impacts on some behaviors than others. Specifically, either scaling approach may be adequate for the measurement of behaviors not clearly falling within the state/event dichotomy.

This study provides an important contribution to the literature on the instrumentation of DBR-SIS scales. Particularly, the inclusion of percentage of time anchors may be more important for some behaviors than others, aligning to Saudargas and Lentz's (1986) conclusion about

Table 2. Results of Nonparametric Tests and 95% Bootstrapped CIs: SDO Error Scores.

Behavior	Level	Group	Mann–Whitney *U*	*p*	Mean rank	*Mdn* $\lvert DBR_{Part} - SDO \rvert$	Bootstrapped 95% CI $\lvert DBR_{Part} - SDO \rvert$ Lower	Upper
Academically engaged	Low	Percent	1,185.50	.002*	51.02	0.70	0.70	1.30
		No percent			70.45	1.70	1.30	2.30
	Medium	Percent	1,496.50	.157	64.12	2.20	1.80	3.20
		No percent			55.21	2.20	1.20	2.20
	High	Percent	1,279.00	.004*	52.48	0.20	0.20	0.20
		No percent			68.75	0.80	0.20	1.80
Disruptive	Low	Percent	1,658.50	.584	61.59	1.70	0.70	2.30
		No percent			58.15	1.30	0.70	1.70
	Medium	Percent	1,622.50	.455	62.15	2.30	2.30	2.30
		No percent			57.50	2.30	1.30	2.30
	High	Percent	1,758.50	.993	60.02	1.50	0.50	1.50
		No percent			59.97	1.50	0.50	1.50
Respectful	Low	Percent	1,688.50	.695	58.88	0.70	0.70	1.30
		No percent			61.30	1.30	0.70	1.30
	Medium	Percent	1,384.50	.043	65.87	3.30	2.30	4.30
		No percent			53.17	2.30	2.30	3.30
	High	Percent	1,652.50	.557	61.68	1.00	1.00	2.00
		No percent			58.05	1.00	1.00	2.00

Note. SDO = systematic direct observation; DBR = Direct Behavior Rating; CI = confidence interval.

Table 3. Results of Nonparametric Tests and 95% Bootstrapped CIs: Expert Error Scores.

Behavior	Level	Group	Mann–Whitney *U*	*p*	Mean rank	*Mdn* $\lvert DBR_{Part} - DBR_{Exp} \rvert$	Bootstrapped 95% CI $\lvert DBR_{Part} - DBR_{Exp} \rvert$ Lower	Upper
Academically engaged	Low	Percent	889.50	.000*	46.40	1.00	0.00	1.00
		No percent			75.83	2.00	2.00	3.00
	Medium	Percent	1,608.00	.411	57.24	3.00	2.00	3.00
		No percent			62.38	2.00	2.00	3.00
	High	Percent	1,279.00	.004*	52.48	0.00	0.00	0.00
		No percent			68.75	1.00	0.00	2.00
Disruptive	Low	Percent	1,657.50	.575	61.60	2.00	1.00	2.00
		No percent			58.14	1.00	1.00	2.00
	Medium	Percent	1,576.50	.305	62.87	1.00	1.00	2.00
		No percent			56.66	1.00	1.00	2.00
	High	Percent	1,595.50	.319	57.43	1.00	1.00	1.00
		No percent			62.99	1.00	1.00	1.00
Respectful	Low	Percent	1,411.50	.049	54.55	1.00	0.00	1.00
		No percent			66.34	1.00	1.00	1.00
	Medium	Percent	1,726.00	.852	60.53	2.00	1.00	2.00
		No percent			59.38	1.00	1.00	2.00
	High	Percent	1,652.50	.557	61.68	1.00	1.00	2.00
		No percent			58.05	1.00	1.00	2.00

Note. DBR = direct behavior rating; CI = confidence interval.

the differences in measuring state versus event behaviors. Therefore, the scaling utilized for behaviors is a particularly relevant consideration when practitioners use DBR-SIS or DBR-like tools. That is, given the genesis of DBR tools and a history of user-created measures, consideration needs to be given regarding the construction of the scale and scaling

of specific behaviors to maximize the desirable psychometric properties of these instruments.

With regard to the effect size estimates, moderate to large effects were observed for those contrasts yielding significant differences between groups. In terms of data-based decision making, it is plausible that such differences could have modest impacts on goal setting and determinations made regarding goal attainment. This is important if the goal is to enhance the psychometric adequacy of rating data (Saal et al., 1980). Consistent with previous research, results from this study suggest that DBR-SIS may be a flexible tool with regard to instrumentation design, while also suggesting that decisions regarding scale construction may need to be considered within the context of specific behaviors (Briesch et al., 2012; Riley-Tillman et al., 2011; Skinner et al., 2000).

Limitations and Future Research

Several limitations are acknowledged in the context of the present investigation. First, the participants represent a sample of convenience. Therefore, the extent to which undergraduate student ratings of behavior relate to teacher ratings remains unknown. Similarly, there are limitations regarding the generalizability of findings; the ratings were obtained in a controlled setting using only brief video clips of student behavior. Further research is needed in in vivo settings with teachers as raters. Second, group equivalence could not be evaluated, due to demographic data being collected on the potential subject pool, rather than actual participants. Third, as previously discussed, no gold standard exists as a criterion for DBR-SIS ratings. Consequently, the criterion scores used (SDO and expert scores) represent only two alternatives to evaluate rating accuracy. Notably, both of these approaches measured duration specifically as the dimension of interest, and consequently accuracy was defined in relation to these estimates of duration.

Additional research is needed to further understand how raters use the DBR-SIS scale to assign ratings to students. In particular, raters are instructed to mentally estimate the duration of each of the target behaviors, but it seems plausible that teachers could factor other dimensions of behavior (e.g., intensity) into their ratings. Therefore, gaining further insight regarding the process by which raters assign ratings and the extent to which they focus on particular dimensions of behavior in assigning ratings would be highly informative.

Conclusion

Given DBR's intended use for data-based decision making, it is important to understand factors that influence rating accuracy. Thus, calls have been made for further investigations regarding the development and evaluation of DBR instrumentation (Christ & Boice, 2009). This study provides preliminary data regarding how the instrumentation of DBR-SIS may affect rating accuracy. Specifically, the inclusion of percentage of time anchors and references to duration were associated with more accurate ratings of AE behavior. In light of these findings, we recommend that behavioral assessment researchers and developers continue to attend to scaling formats utilized and study potential impacts on rating accuracy.

Authors' Note

Opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.

Acknowledgments

The authors would like to thank Shannon Brooks, Rose Jaffery, and Rohini Sen for their assistance with this study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this article was supported by grants provided by the Institute for Education Sciences, U.S. Department of Education (R324B060014, R324A110017).

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*, 227–267.

Briesch, A. M., Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2012). The influence of alternative scale formats on the generalizability of data obtained from Direct Behavior Rating Single-Item Scales (DBR-SIS). *Assessment for Effective Intervention*, *38*, 127–133. doi:10.1177/1534508412441966

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

Chafouleas, S. M., Christ, T., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. (2007). Generalizability and dependability of Direct Behavior Ratings to measure social behavior of preschoolers. *School Psychology Review*, *36*, 63–79.

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

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

Christ, T. J., & Boice, C. (2009). A brief review of nomenclature, components, and formatting to inform the development of Direct Behavior Rating (DBR). *Assessment for Effective Intervention*, *34*, 242–250.

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.

Diener, E., Smith, H., & Fujita, F. (1995). The personality structure of affect. *Journal of Personality and Social Psychology*, *50*, 130–141.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, *79*, 314–316.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum.

Hartley, J., Trueman, A., & Rodgers, A. (1984). The effects of verbal and numerical quantifiers on questionnaire responses. *Applied Ergonomics*, *15*, 149–155.

Jaffery, R., Johnson, A. H., Bowler, M. C., Riley-Tillman, T. C., Chafouleas, S. M., & Harrison, S. E. (2015). Using consensus building procedures with expert raters to establish true score estimates of behavior for direct behavior rating. *Assessment for Effective Intervention*, *40*, 195–204.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*, 110–129.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 1–55.

Lucas, R. E., Diener, E., & Larsen, R. J. (2009). Measuring positive emotions. In E. Diener (Ed.), *Assessing well-being* (pp. 139–155). Dordrecht, The Netherlands: Springer.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement*, *49*, 33–43. doi:10.1177/0013164489491004.

Nussbaum, E. M. (2014). *Categorical and nonparametric data analysis: Choosing the best statistical techniqu*e. Florence, KY: Taylor & Francis.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903.

Riley-Tillman, T. C., Christ, T. J., Chafouleas, S. M., Boice-Mallach, C. H., & Briesch, A. (2011). The impact of observation duration on the accuracy of data obtained from Direct Behavior Rating (DBR). *Journal of Positive Behavior Interventions*, *13*, 119–128. doi:10.1177/1098300710361954

Rohrmann, B. (2007). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data* (Project Report for the University of Melbourne). Retrieved from http://rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Saudargas, R., & Lentz, F. (1986). Estimating percent of time and rate via direct observation: A suggested observational procedure and format. *School Psychology Review*, *15*, 36–48.

Skinner, C. H., Rhymer, K. N., & McDaniel, E. C. (2000). Naturalistic direct observation in educational settings. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Conducting school-based assessments of child and adolescent behavior* (pp. 21–54). New York, NY: Guilford Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Tapp, J. (2002). Multiple Option Observation System for Experimental Studies (MOOSES) [Software]. Retrieved from http://mooses.vueinnovations.com/overview/mooses-overview

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.