Project VIABLE: Investigating the Influence of Scale Characteristics on Direct Behavior Rating Data

Albee Ongsuco[1], Ajlana Music[1], Amy M. Briesch[2], Stephen P. Kilgus[2], Sandra M. Chafouleas[2], T. Chris Riley-Tillman[1], & Theodore J. Christ[3]

East Carolina University[1] University of Connecticut[2] University of Minnesota[3]

Introduction

Direct Behavior Rating (DBR) represents a flexible, usable, and defensible means by which to assess student behavior (Chafouleas, Riley-Tillman, & Sugai, 2007). Emerging research relative to DBR has suggested use to be potentially viable across several assessment purposes, including screening (Chafouleas, Kilgus, & Hernandez, in press) and progress monitoring (Chafouleas, McDougal, Riley-Tillman, Panahon, & Hilt, 2005; Chafouleas, Riley-Tillman, Sassu, LaFrance, & Patwa, 2007). Despite emerging supporting evidence, there remains a lack of consensus regarding which scaling features should be utilized when constructing the DBR scale. For instance, Chafouleas and colleagues (2005), and Chafouleas, Riley-Tillman, and colleagues (2007) examined a DBR scale that included a graphical line as a visual cue, divided into 6 equal gradients (labeled 0-5), with percentage anchors at each gradient (e.g., 0%, 20%, 40%). In contrast, Chafouleas, Christ, and colleagues (2007) utilized a 105mm DBR scale, separated into 11 equal gradients (0-10), with percentage anchors at the beginning (0%), middle (50%), and end (100%) of the scale.

Empirical work in related literature bases has suggested that variations among scales similar to DBR may not differentially affect the reliability and validity of the resulting data (Bendig, 1954; Jacoby & Matell, 1971; Matell & Jacoby, 1971). Yet, it remains to be seen whether these findings extend to the use of DBR. As such, the purpose of the current study was to evaluate several scaling arrangements in an attempt to specify the scale which may facilitate the most technically adequate ratings of student behavior. Through this work, we intend to specify a standard DBR scale that may be used within future investigations. Recommended standard instrumentation, such as scale, will serve to unify both past and future psychometric investigations of DBR, as well as inform applied use of the measure.

Method

Participants included 198 undergraduate students enrolled in an introductory psychology course at a large Southeastern university. The majority (58%) of participants were female, and identified themselves as either Caucasian (61%) or African American (24%). Participants were randomly assigned to a condition upon entrance to the experimental settings (see Table 1). Enrollment determined which graphical presentation of the DBR scale the participant used to rate four target students in terms of disruptive behavior and academic engagement at the end of each of four 3-minute clips. Specifically, the DBR scale used by participants varied across three areas, including: (a) type of scale (i.e., continuous line [0-100%] or discrete scale [1-10]), (b) number of scale gradients provided (i.e., 5 or 10), and (c) length of scale (50 or 100mm).

Table 1. *Description of Experimental Condition*.

	No Line		Line	
	5 gradients	10 gradients	5 gradients	10 gradients
50 mm	Group 1	Group 2	Group 5	Group 6
100 mm	Group 3	Group 4	Group 7	Group 8

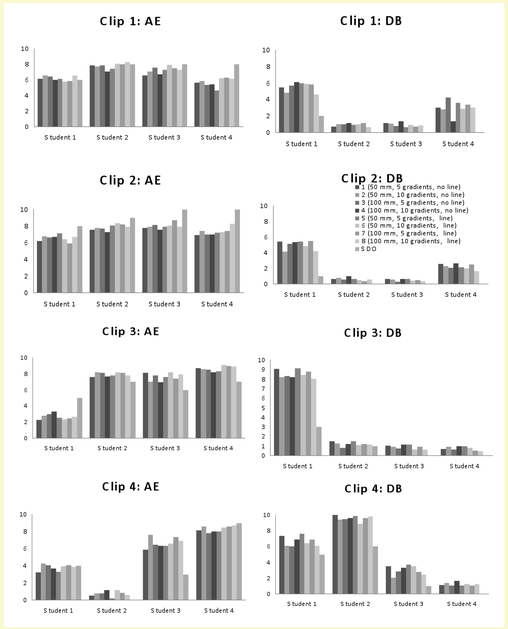
Preparation of this poster was supported by a grant from the Institute for Education Sciences (IES), U.S. Department of Education (R324B060014). For additional information, please direct all correspondence to Sandra Chafouleas at sandra.chafouleas@uconn.edu

Results

Table 2. *Results of Generalizability Studies Within Condition*

	5 gradients AE	5 gradients DB	10 gradients AE	10 gradients DB
No Line				
50 mm	Rater = 5.0 Occasion = 2.0 Person = 5.3 Rater*Occ = 3.9 Rater * Person = 6.9 Occ * Person = 44.9 Error = 31.9	Rater = 2.7 Occasion = 4.0 Person = 30.7 Rater*Occ = 1.2 Rater * Person = 1.9 Occ * Person = 42.1 Error = 17.4	Rater = 10.3 Occasion = 0 Person = 1.5 Rater*Occ =1.5 Rater * Person = 5.4 Occ * Person = 45.8 Error = 35.7	Rater = 5.2 Occasion = 0.7 Person = 24.5 Rater*Occ = 2.0 Rater * Person = 2.8 Occ * Person = 43.8 Error = 20.9
100 mm	Rater = 14.5 Occasion = 2.2 Person = 0.7 Rater*Occ = 3.1 Rater * Person = 5.9 Occ * Person = 39.6 Error = 33.9	Rater = 4.1 Occasion = 0 Person = 24.7 Rater*Occ = 1.4 Rater * Person = 2.6 Occ * Person = 43.7 Error = 23.5	Rater = 14.5 Occasion = 1.5 Person = 2.1 Rater*Occ = 3.1 Rater * Person = 7.4 Occ * Person =38.6 Error = 32.9	Rater = 4.9 Occasion = 3.0 Person = 27.5 Rater*Occ = 1.7 Rater * Person = 4.6 Occ * Person = 39.4 Error = 18.9
Line				
50 mm	Rater = 8.5 Occasion = 2.5 Person = 0 Rater*Occ = 4.9 Rater * Person = 8.6 Occ * Person = 44.2 Error = 31.2	Rater = 2.1 Occasion = 3.7 Person = 31.0 Rater*Occ = 4.4 Rater*Person = 3.3 Occ * Person = 37.4 Error = 18.2	Rater = 6.7 Occasion = 0 Person = 9.2 Rater*Occ = 3.4 Rater*Person = 13.1 Occ * Person = 38.6 Error = 29.0	Rater = 7.0 Occasion = 3.3 Person = 28.8 Rater*Occ = 1.9 Rater * Person = 5.9 Occ * Person = 32.2 Error = 20.8
100 mm	Rater = 6.6 Occasion = 0 Person = 8.9 Rater*Occ = 4.3 Rater * Person = 8.3 Occ * Person = 43.2 Error = 28.7	Rater = 2.5 Occasion = 0 Person = 32.4 Rater*Occ = 2.6 Rater * Person = 3.2 Occ * Person = 40.7 Error = 18.6	Rater = 7.5 Occasion = 1.3 Person = 6.8 Rater*Occ = 5.8 Rater * Person = 7.7 Occ * Person = 42.1 Error = 28.8	Rater = 1.8 Occasion = 4.0 Person = 22.2 Rater*Occ = 1.6 Rater * Person = 3.9 Occ * Person = 47.0 Error = 19.6

Figure 1. *Mean DBR and SDO Ratings for Each Target Student*



All DBR ratings were converted and expressed using an 11-point (i.e. 0-10) discrete scale in order to allow for meaningful comparisons across conditions. Of the 32 ratings conducted (i.e. 4 students rated on 2 behaviors across 4 clips), participant ratings fell within 1 point of the true score (as determined by systematic direct observation, using momentary time sampling of 1-second intervals) 39% of the time when rating academic engagement, and 36% of the time when rating disruptive behavior. When the margin of error was slightly expanded, participant ratings fell within 2 points of the true score 71% of the time when rating academic engagement and 60% of the time when rating disruptive behavior. A one-way ANOVA examined the relationship between the type of DBR scale utilized and the accuracy with which participants were able to rate a target student on a particular behavior, after removing all other rater biases and errors (i.e. differential accuracy). Overall, results were not found to be statistically significant, $F(7, 192) = 1.19$, $p = .31$, $\eta^2 = .04$. This suggests that rating accuracy did not vary substantially depending on the scale that participants were given.

Supplemental analyses were conducted using Generalizability Theory (GT; Brennan, 2003), in order to determine (1) the proportion of variance attributable to identified facets, and (2) the number of DBR ratings needed in order to derive a dependable estimate of student behavior. Both generalizability (G) and dependability (D) studies were conducted within experimental condition using a fully-crossed, random effects model for person (p) by rater (r) by observation (o). Results of the G studies varied widely, depending on the behavior and type of scale analyzed (see Table 2). Results for both behaviors were similar in that large variance components were identified for the interaction between persons and occasions. In contrast, however, the proportion of variance attributable to persons was much higher in the case of disruptive behavior than in the case of academic engagement. Subsequent D studies were conducted in order to determine the number of DBR ratings needed in order to derive a dependable estimate of student behavior. Again, estimates varied widely across behaviors. Given the same initial model, Dependability (Φ) coefficients for Academic Engagement ranged from 0 to .46, whereas the coefficients for Disruptive Behavior ranged from .63 to .75.

Summary and Conclusions

Results indicated that rating accuracy did not vary significantly across DBR scales. When ratings were converted to an 11-point discrete scale, the largest mean discrepancy between different DBR scales was 1.7 points for Academic Engagement and 2.89 points for Disruptive Behavior (see Figure 1). In fact, the mean DBR rating for a given student most often fell within the same scale gradient across conditions. Although no clear patterns emerged with regard to the rating of academic engagement, participants were more likely to over-rate than to under-rate disruptive behavior. Results of generalizability analyses varied widely depending on the behavior and scale type examined. First, a greater proportion of unexplained error variance was noted for ratings of academic engagement (28.7-35.7%) than for ratings of disruptive behavior (17.4-23.5%). Second, the proportion of variance attributable to the target of measurement (i.e. person) was much larger when measuring disruptive behavior than academic engagement. In other words, is was easier to capture differences between students when measuring disruptive behavior. One interesting finding was the fact that the greatest proportion of variance in ratings of both behaviors was attributable to the person by occasion interaction. This indicates that student behavior varied widely from one observation occasion to the next. Such a finding lends support to the validity of DBR as a tool for monitoring student behavior, particularly relative to its sensitivity to change.

Overall, results support one of the commonly stated features of DBR; namely, that the method may be flexibly constructed (Chafouleas, Riley-Tillman, & Sugai, 2007). Yet, we raise the caveat that one should not assume that past, present, or future DBR-related psychometric findings generalize to any measure beyond that which was investigated within any particular study. Although accuracy was not found to vary across scales in the current study, such results may not be guaranteed in future investigations.