Project VIABLE

University of Connecticut

East Carolina University

University of Minnesota

Direct Behavior Rating: Impact of Behavioral Wording on Data Accuracy

Rose Jaffery[1], Rohini Sen[1], Ajlana Music[2],
Sandra M. Chafouleas[1], T. Chris Riley-Tillman[2], Theodore J. Christ[3]

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

Introduction

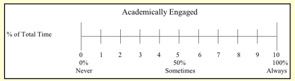
Direct Behavior Rating (DBR) is a form of behavioral assessment that has the potential to have comparable defensibility to data obtained through the use of systematic direct observation (SDO; Chafouleas, Riley-Tillman, & Christ, 2009). However, there have been few empirical investigations regarding the selection of target behaviors and impact of target behavior wording (positive/negative) on the accuracy of DBR data. Results of a preliminary study (Riley-Tillman et al., 2009) suggested that individuals are able to produce more accurate ratings when asked to judge global rather than specific behaviors, however results were inconsistent with regard to target wording (positive/negative). More specifically, results suggested the use of positive wording when rating *academic engagement* (AE); however, either positive or negative wording appeared to be similarly acceptable for *disruptive behavior* (DB). A follow-up study by Christ et al. (2010) examining global behaviors found that behavior connotation (positive/negative) did not have a substantial effect on rating accuracy for either AE or DB. However, raters more accurately rated some behaviors over others, indicating that connotative wording might influence accuracy of DBR data for some, but not all behaviors.

The purpose of this study was to extend previous work regarding use of global behaviors to evaluate whether data accuracy is impacted by (a) connotative wording of the target behavior (e.g., disruptive vs. non-disruptive) and (b) the level at which the target student displayed that behavior. It was hypothesized that, as in previous findings, accuracy would be minimally influenced by wording for AE and DB, and that there will be more substantial rater error and bias for *respectful* (RS) given greater challenge establishing a universal operational definition.

Method

Materials. Video footage of elementary school students was recorded during simulated classroom instruction and cut into nine 1-minute clips. Video clips were purposefully selected to reflect varying levels (low, medium, high) of AE, DB, and RS displayed by two target students. Clips were then randomly ordered, and DBR forms created to match. Two Rating Packets resulted – one with positive (i.e., academically engaged, respectful, non-disruptive) and the second with negative (i.e., academically unengaged, disrespectful, disruptive) wording.

Figure 1. *Example DBR Single-Item Scale*



Participants. Participants included 113 undergraduate students enrolled in an introductory psychology course at a large university in the southeastern United States. Six study sessions were conducted - three received the positive condition and three received the negative.

Procedures. First, participants viewed a video describing DBR and how it can be used to assess student behavior. The three target behaviors (either positively or negatively worded) were also included in the video with explicit definitions and examples. Next, participants in both conditions viewed the same nine 1-minute video clips of an elementary-level classroom and rated the target student's target behavior (AE, RS, or DB) for each clip using a Rating Packet with all positively worded or negatively worded behaviors. Prior to each clip, the student and behavior to be rated was introduced, then the 1-minute video clip was viewed. The Rating Packets consisted of DBR scales from 0-10 for each clip. After each clip participants were instructed to estimate the percentage of time the target student displayed the target behavior and to mark it on the corresponding DBR scale (0=0%, 10=100%).

Two criterion measures were used to evaluate the accuracy of data. Rating scores based on SDO using momentary time sampling procedures with 1-second intervals were obtained for each of the nine clips. Expert DBR ratings for each clip were also obtained using a consensus-building procedure with individuals highly familiar with assessing student behavior using DBR.

Results

Data Analysis. Differences and absolute differences between mean participant DBR data and Expert DBR data (DBR_{Part}-DBR_{Exp} and |DBR_{Part}-DBR_{Exp}|), as well as between DBR and SDO data (DBR_{Part}-SDO and |DBR_{Part}-SDO|) were calculated (see Table 1). Difference scores (i.e., DBR_{Part}-SDO and DBR_{Part}-DBR_{Exp}) indicate the tendency of participants to *under-* or *over*-estimate data when compared to each criterion (i.e., *rater bias*). Absolute difference scores (i.e., |DBR_{Part}-SDO| and |DBR_{Part}-DBR_{Exp}|) indicate the general magnitude of the difference between participant ratings and the criterion scores (i.e., *rater error*). Rater bias and rater error scores close to zero are preferable as they indicate ratings closer to the criterion, thus more accuracy.

Three 2 (wording: positive, negative) x 3 (level: low, medium, high) between-subjects ANOVAs were conducted for each behavior (AE, DB, and RS) to analyze the impact of connotative wording and presentation level of behavior on participants' ratings of each target behavior. Post hoc comparisons using between-subjects t-tests were also conducted to compare mean rater bias and error values for positive versus negative conditions within each of the three levels for all three behaviors (see Table 2).

Table 1. *Descriptive Statistics for SDO, Participant DBR (DBR_{Part}), and Expert DBR (DBR_{Exp}) scores*

			SDO		DBR_{Part}		DBR_{Exp}	
Behavior	Wording	Level	*Mean*	*SD*	*Mean*	*SD*	*Median*	*SD*
Academically Engaged	Positive	Low	1.7	0	1.44	1.38	1.0	0
		Medium	4.8	0	5.44	2.42	4.0	0
		High	9.8	0	9.07	1.85	10.0	0
Academically Unengaged	Negative	Low	8.3	0	7.91	2.2	9.0	0
		Medium	5.2	0	4.14	2.64	6.0	0
		High	0.2	0	0.64	1.86	0.0	0
Non-Disruptive	Positive	Low	8.3	0	6.35	2.83	8.0	0
		Medium	4.3	0	2.62	2.01	3.0	0
		High	1.5	0	1.75	3.02	1.0	0
Disruptive	Negative	Low	1.7	0	2.47	2.05	2.0	0
		Medium	5.7	0	7.09	1.56	7.0	0
		High	8.5	0	9.09	1.41	9.0	0
Respectful	Positive	Low	1.7	0	1.82	2.36	1.0	0
		Medium	5.3	0	1.56	2.11	3.0	0
		High	10.0	0	7.04	2.36	10.0	0
Disrespectful	Negative	Low	8.3	0	7.97	2.03	9.0	0
		Medium	4.7	0	7.69	1.99	7.0	0
		High	0.0	0	1.29	1.87	0.0	0

Table 2. *Comparisons of Rater Bias and Error for Positive and Negative Wording Conditions within Levels using T-Test*

			Expert Criterion Score				SDO Criterion Score			
			Rater Bias		Rater Error		Rater Bias		Rater Error	
Behavior	Level	Wording	Mean Difference	t	Mean Difference	t	Mean Difference	t	Mean Differ ence	t
Academically Engaged	Low	Positive vs. Negative	-0.15	-3.10**	0.03	0.62	0.33	4.36**	0.02	0.55
	Medium	Positive vs. Negative	-3.29	-6.91**	0.49	1.52	-1.69	-3.57**	0.46	1.59
	High	Positive vs. Negative	-0.18	-1.90	-0.05	-0.91	0.05	0.24	-0.05	-1.22
Disruptive	Low	Positive vs. Negative	2.12	4.58**	-0.03	-0.60	2.72	5.87**	-0.05	-1.00
	Medium	Positive vs. Negative	0.47	1.39	-0.43	-2.37*	3.07	9.08**	-0.65	-2.30**
	High	Positive vs. Negative	-0.18	-2.80**	-0.05	-2.06*	0.06	0.70	-0.05	-2.11*
Respectful	Low	Positive vs. Negative	-0.07	-1.34	-0.05	-0.99	0.28	3.45**	-0.06	-1.52
	Medium	Positive vs. Negative	2.13	5.51**	-0.37	-1.62	6.72	17.44**	-0.83	-3.25**
	High	Positive vs. Negative	-0.44	-8.72**	-0.26	-4.75**	-0.27	-4.89**	-0.26	-4.76**

*significant at 0.05, ** significant at 0.01

Overall impact of level and wording on accuracy as indicated by between-subjects ANOVAs:

Respect

- SDO results showed significant interaction effect of level and wording condition on rater bias *F* (265.72, 2) as well as rater error *F*(6.93, 2).
- Expert DBR results showed significant interaction effect of level and wording condition on rater bias (*F*[15.16, 2]), but no statistically significant interaction effect of level and wording on rater error. For rater error, the main effects of wording and level were statistically significant (*F*[8.08, 1] and *F* [186.16, 2], respectively).

Academic Engagement

- For SDO, the interaction effect of level and wording for rater bias (*F*[10.96,2]) is statistically significant while the interaction effect is not statistically significant for rater error (*F*[2.65, 2]). For rater error, the main effect of wording is not statistically significant (*F*[2.14, 1]), but level is significant (*F*[290.10,2]).
- For Expert DBR, the interaction effect on rater bias (*F*[32.93,2]) is statistically significant while the interaction effect of level and wording is not statistically significant for rater error (*F*[2.33, 2]). For rater error, wording is not statistically significant (*F*[2.00,1]) but level is significant (*F*[311.17,2]).

Disruption

- SDO results show that for disruptive behavior, there is a significant interaction effect of wording and level on both rater bias as well as rater error (*F*[15.41, 2] and *F*[7.35, 2], respectively).
- Expert DBR results show that for disruptive behavior, there is significant interaction effect of wording and level on both rater bias as well as rater error (*F*[9.43,2] and *F*[4.23,2], respectively).

Overall, results show that participants' DBR data corresponded fairly well with either criterion.

- As expected, across all three behaviors, *medium levels* of behavior resulted in *reduced accuracy*.
- In addition, ratings of RS resulted in the largest difference scores for either wording, indicating *reduced accuracy.*

More in depth comparisons allowed us to evaluate specific inaccuracies (Table 2). Results of between-subjects t-tests looking at comparisons of rater bias and error for positive and negative wording conditions within level indicate that overall:

- For *academically engaged* behavior there is a slight advantage for *positive wording* for all comparisons regardless of criterion score.
- For *disruptive* behavior, there is a slight advantage for *negative* wording when expert criterion score is used. However, when SDO criterion score is used, there is an advantage for *positive* wording.
- For *respectful* behavior, t-test results indicate that overall *negative* wording may have an advantage over positive wording, except in the case of low level of respectful behavior.

Summary and Conclusions

Overall, results indicate that with minimal training, participants' ratings corresponded fairly well with both SDO and expert DBR scores. This is consistent with previous findings and contributes to the defensibility of DBR as a method for assessing student behavior that can collect reliably accurate data (Christ et al., 2011; Riley-Tillman et al., 2009). However, results also indicate that the connotative wording of behavioral targets and level of behavior displayed in the sample can impact the accuracy of DBR data.

In general, findings show that *positive* wording for AE is preferable as it resulted in more accurate data overall (i.e., *academically engaged* is preferable vs. academically unengaged). *Negative* wording for DB and RS is preferable (i.e., *disruptive* is preferable vs. non-disruptive, and *disrespectful* is preferable vs. respectful). In terms of the impact of level on ratings, video clips that displayed behaviors at a *medium* level resulted in more error/bias. Across all behaviors, overall ratings for RS indicated much worse accuracy. However, expert DBR and SDO scores for RS often did not correspond, indicating that there may be characteristics unique to RS impacting how it is scored when using different criterions. Future research should continue to explore RS as a behavioral target and focus on replicating findings from this highly-controlled study in a practical setting to determine if the effects maintain across settings.

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