Direct Behavior Rating: Impact of Behavioral Wording on Data Accuracy

Rose Jaffery¹, Rohini Sen¹, Ajlana Music², Sandra M. Chafouleas¹, T. Chris Riley-Tillman², Theodore J. Christ³
University of Connecticut¹, University of Minnesota², East Carolina University³

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, 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 purposively selected to reflect varying levels (low, medium, high) of AE, DB, and RS displayed by two target students. Clips were then randomly ordered, and participants were matched. 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.

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 condition.

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 on the corresponding DBR scale (0-9, 10-100%).

For Expert DBR, the interaction effect on rater bias (F[15.16, 2]) is statistically significant while the interaction effect of wording and level on both rater bias as well as rater error (F[15.41, 2]) and for rater error (F[15.43, 2] and F[4.31, 2], respectively).

Overall results show that participants’ DBR data corresponded fairly well with either condition. 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., 2013; 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 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 DB 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.