Comparing the Psychometric Properties of Behavior Assessment Methods: Systematic Direct Observation & Direct Behavior Rating

Amy M. Briesch1, Sandra M. Chafouleas2 & T. Chris Riley-Tillman3
Northeastern University1 University of Connecticut2 East Carolina University3

Statement of the Problem

• Given increased demand for accountability and data-based decision making in schools, the search is underway to identify appropriate measurement tools for social behavior.
• Although systematic direct observation (SDO) has typically been regarded as the method of choice when formatively assessing social behavior within research contexts, there exists some concern that the practical utility of SDO data may be limited by its scope in applied settings such as schools. Extant research findings suggest that the sheer number of systematic observations required in order to draw a generalizable picture of student behavior may have the potential to overwhelm existing systems and resources (e.g., Hintze & Matthews, 2004).
• Use of direct behavior rating (DBR) may provide a more feasible alternative in terms of frequent data collection; however, some concern has been raised regarding the influence of rater effects (e.g., halo error) on the reliability and accuracy of data. Furthermore, although early evidence appears promising, research regarding the psychometric properties of DBR is still in its infancy.

Method

Data collected over the course of 10 consecutive school days during a 45-minute group lesson (phonemic awareness, phonics, mathematics) (Data collection videotaped and subsequently coded using momentary-time sampling procedures by 2 trained observers.

This study was supported by a grant from the Institute for Education Sciences, U.S. Department of Education (R324B000014). For additional information, please direct all correspondence to Amy Briesch at a.briesch@neu.edu


Results and Discussion

Rater effects
• SDO: Both the rater facet (0%) and the persons by rater interaction (0%) contributed negligible variance to the model, suggesting that neither interobserver agreement nor rater bias were of concern.
• DBR: One-quarter of the variance in DBR data was attributable to rater-related effects.

• The fact that overall rating differences between the two teachers were identified was not surprising, given descriptive differences noted in teachers’ use of the DBR scale (i.e., Mean rating for Teacher 1 = 90.89%, Mean rating for Teacher 2 = 80.32%).
• The variance component for the interaction between persons and raters accounted for 20% of rating variance, indicating that teachers varied in their perceptions of the relative standing of particular students (e.g., rater bias effect).
• These findings suggest that recordings must either be analyzed within raters, thereby focusing on changes in individuals of student perception of the problem, or teachers must engage in training in order to minimize reliance on global perception and move closer in line with objective observation.

Generalizability study results
• Nearly 50% (47 teachers, 48 researchers) of the variance in scores was attributable to the facet of person, suggesting that both methods were equally sensitive to inter-individual differences in academic engagement.
• The remaining 50% of variance in scores, however, was accounted for in meaningfully different ways depending on the assessment method, thus suggesting different implications for the use of either an SDO or DBR assessment approach.

Table 1

<table>
<thead>
<tr>
<th>Type of Rating</th>
<th>Generalizability Study Results</th>
<th>Socioeconomic Status (SES) Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

Results of the current investigation suggest important considerations relevant to the selection of behavior assessment methods across varying assessment purposes. It should not come as a surprise that more dependable estimates of student engagement were obtained using SDO than DBR, given both the precision of the recording procedure (i.e., behavior recorded every 15 seconds versus once every 15 minutes) and the nature of the rating task (i.e., analogue observation of videotaped behavior versus in vivo rating during regular teaching activities). However, the selection of appropriate assessment tools is rarely based on numbers alone, requiring simultaneous consideration of issues related to defensibility and feasibility. Results of the current study, for example, suggest that a level of reliability sufficient for the purposes of rank-order screening (i.e., .70) may be obtained given the collection of 3 SDO or 20 DBR data points. Although the choice may seem obvious at a surface level, several factors may intermix to determine selection. For example, in situations in which decisions need to be made quickly, use of SDO would be the most appropriate choice given that fewer data points would be needed. When time is not a limiting factor, however, school psychologists may find teachers’ daily use DBR over the course of a few weeks to be less intrusive (to both the natural classroom ecology and the school psychologist’s schedule) than conducting 45-60 minutes worth of targeted observation. Furthermore, of particular importance for those situations in which behaviors occur infrequently or teacher perception of the target behavior is deemed an important outcome variable, it appears that the collection of a sufficient amount of teacher-generated data may serve as a viable alternative to scheduled observations.