Teaching and Teacher Education 59 (2016) 414–419

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Behaviorally anchored rating scales: An application for evaluating teaching practice

Michelle Martin-Raugh*, Richard J. Tannenbaum, Cynthia M. Tocci, Clyde Reese

Educational Testing Service (ETS), 660 Rosedale Rd., Princeton, NJ, 08541, USA

HIGHLIGHTS

- We compared Behaviorally Anchored Rating Scales to the Framework for Teaching.
- BARS provide behavioral anchors delineating levels of performance.
- Nineteen raters, users of the FfT trained to use BARS, evaluated teacher lessons.
- We report rater agreement, usability judgments, and preferences for both tools.
- Raters, although familiar with the FfT, reported favorable reactions to BARS.

ARTICLE INFO

Article history:
Received 25 January 2016
Received in revised form
18 July 2016
Accepted 22 July 2016

Keywords:
Behaviorally anchored rating scales (BARS)
Framework for Teaching (FfT)
Appraisal
Evaluation

ABSTRACT

We developed Behaviorally Anchored Rating Scales (BARS) for measuring teaching practice, and compared them to the well-established Framework for Teaching (FfT; Danielson, 2013). BARS provide behavioral anchors delineating levels of performance via a set of behaviors. Our BARS focused on two dimensions of teaching, leading a classroom discussion and making content and practices explicit. We examined how a) rater agreement for BARS compares to rater agreement using the FfT, and b) how BARS and the FfT compare regarding perceived ease of use, perceived accuracy, and perceived advantages and disadvantages. Nineteen raters, who are users of the FfT and were trained to use BARS, independently evaluated video-taped teacher lessons using both methods. Rater agreement was higher for the FfT, which may, in part, be a factor of the raters' greater experience with that rating system. Nonetheless, raters reported that there are many aspects of BARS that are desirable.

There is a great deal of agreement among both researchers and educators that teachers have a large effect on the lives of elementary school students (Harris & Rutledge, 2010; McCartney, Dearing, Taylor, & Bub, 2007), and that effectively measuring teaching performance is an important area of inquiry. Recent education reforms aiming to improve student performance have focused, in part, on improving teacher selection, preparation, and evaluation (Goe, 2007). However, there is not yet consensus among educational researchers about the specific indicators that define quality teaching nor, not surprisingly therefore, about the best ways to measure teaching practice (cf. Ball & Hill, 2008). Prior research suggests that traditional principal evaluations of teachers insufficiently differentiate between more and less effective teachers and provide an inadequate foundation for highlighting teacher needs for training and improvement (Danielson, 1996; Medley & Coker, 1987). More rigorous teacher evaluation tools may inform teacher development at crucial junctures, such as certification and selection, and may be used to shape educator training and professional development (Glazerman et al., 2010; Jamil, Sabol, Hamre, & Pianta, 2015; Bill and Melinda Gates Foundation, 2010).

In this study, we report our efforts to develop and evaluate Behaviorally Anchored Rating Scale (BARS), a type of performance rating scale featuring narrative behavioral anchors at scale points (Smith & Kendall, 1963), for use in measuring observed teaching practice for elementary school teachers teaching Kindergarten through 6th grade. The primary purpose of this study is to evaluate the viability of BARS for evaluating teaching practice. Specifically, we describe the development steps and potential benefits of this measure, and compare some of its properties to those of the commonly used Framework for Teaching (Danielson, 1996; 2007; 2013). Our goal is to consider the possibility of using BARS in the

* Corresponding author.
E-mail address: Mmartin-raugh@ets.org (M. Martin-Raugh).

http://dx.doi.org/10.1016/j.tate.2016.07.026
0742-051X/

service of teacher evaluation, preparation, or development via an exploratory, preliminary study. We have included the Framework for Teaching (FfT) as one reasonable point of reference for facilitating an evaluation of the potential merits of BARS.

1. Assessing teaching quality

There is no one agreed upon definition of teaching quality and effectiveness. One relatively narrow view of teaching quality defines it as the ability to produce gains in student achievement scores on standardized assessments (Little, Goe, & Bell, 2009). However, teachers are not solely responsible for students' learning and test scores, as other factors outside of a teacher's control, such as peers, family members, student abilities, and the school environment affect student learning (Goe, Bell, & Little, 2008; Little et al., 2009). Moreover, standardized assessments are limited in the information they impart, and learning may not be fully captured via scores on assessments (Goe et al., 2008; Little et al., 2009). Thus, an approach to evaluating teaching quality that focuses on specific observable teacher behaviors, likely related to student learning and development, offers a more expanded view of teaching quality (McCloy, 2013).

Ball and colleagues have proposed what they call *high-leverage practices* (HLPs) for teachers, which are "practices at the heart of the work of teaching that are most likely to affect student learning" (Ball & Forzani, 2010, p. 43). One example of a high leverage practice is making content and practices (e.g., specific texts, problems, ideas, theories, processes) explicit to students through explanation, modeling, representations, and examples. Another is effectively leading a group or class discussion. High-leverage practices describe critical characteristics of effective teaching (Ball & Forzani, 2010). However, empirical research on these practices is limited; the current study is one of the first to study their utility as descriptors of teaching to be implemented in evaluating practice. These high leverage practices can be conceptualized as dimensions of effective teaching that may be evaluated via classroom observation.

High-leverage practices are expected to apply across grade levels and subjects. It is reasonable to propose that making content explicit to students, for example, generalizes to all classroom teachers. What may vary by grade level and subject, however, is the specific nature of the evidence that a teacher demonstrates in order to make content explicit; but the fundamental construct remains the same. In this study, we are focusing specifically on elementary school teachers and on English language arts (ELA) and mathematics because the elementary grades serve as the foundation for subsequent learning, and ELA and mathematics are prominent among the essential common core subject areas (e.g., Entwisle & Hayduk, 1988; Porter, McMaken, Hwang, & Yang, 2011).

Many of the current approaches to teaching evaluation for school teachers involve holistic rubrics or performance appraisal instruments, such as the Teacher Performance Evaluation Rubric or Compass rubric. Tools such as these require raters to make an overall judgment about the quality of performance, resulting in scoring that is assumed to be easy, cost-effective and accurate (Jonsson & Svingby, 2007). However, low levels of rater reliability appear to be a persistent issue in this type of observation system (Casabianca, Lockwood, & McCaffrey, 2015). McCaffrey and colleagues (McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2014) reported that correlations among rater errors may substantially distort teacher observation ratings. Moreover, prior research has shown that rater effects accounted for 25%–70% of the variance in observation ratings (Bill and Melinda Gates Foundation, 2012; Casabianca et al., 2013; Hill, Charalambous, & Kraft, 2012).

Rater error contributing to unreliability can take many forms. For instance, raters can differ in the extent to which they make severe or lenient ratings (Kingsbury, 1922; Landy & Farr, 1980). Ratings may also be subject to "halo error," which is a bias resulting in a rater evaluating a behavior based on positive or negative impressions about the individual being assessed (Thorndike, 1920). Additionally, raters may tend to assign scores in the middle of the score range rather than using the full scale, resulting in a central tendency bias (Saal, Downey, & Lahey, 1980). The measurement of teaching quality may be improved through an increased emphasis on teaching behavior and more clearly behaviorally defined scale anchors that are intended to reduce rater biases, and, consequently, result in more reliable and accurate ratings.

Our study explored the potential value of applying such a rating scale to the observation of teaching practice in elementary school. We collected rater judgments about the efficacy of the behaviorally anchored rating scales (BARS) and how its use compared to the FfT, and also compared some basic measurement properties across the different rating approaches.

2. Behaviorally anchored rating scales (BARS)

BARS may afford several advantages over traditional evaluation methods. One advantage stems from the fact that subject matter experts (SMEs) who are familiar with a job and its demands (teachers, in this case) provide information at each step in the development process used to build the scales (Schwab, Heneman, & DeCotiis, 1975). To build BARS, first, critical incidents (Flanagan, 1954) depicting highly effective or ineffective behaviors performed on the job are collected from SMEs. Second, the critical incidents are edited such that redundancies across incidents are removed and they conform to a common format. Third, in a step often referred to as *retranslation* (Schwab et al., 1975), SMEs are asked to evaluate the performance dimension the behavior may be classified into. Fourth, SMEs rate each incident for effectiveness so that the means of these ratings can be used to position the critical incidents as scale anchors. Incidents that do not meet a predetermined criterion of agreement among SMEs are discarded.

The input provided by SMEs at each stage of the development process is likely to result in anchor terminology that is specialized and relevant to the job in question, which may positively impact the reliability of the ratings collected (Schwab et al., 1975). Additionally, the retention of only incidents that reach a high level of expert agreement may reduce central tendency and leniency errors (Smith & Kendall, 1963). Finally, those that are evaluated using BARS may be more likely to react favorably to their evaluations knowing that SMEs with similar backgrounds to their own contributed to the development of the scales (Schwab et al., 1975).

Although the use of ratings to gauge performance is widespread across a variety of professions, educators and researchers alike have had some dissatisfaction with these measures as a result of their vulnerability to intentional and unintentional bias (Landy & Farr, 1980). BARS are intended to reduce bias and subjectivity in rater judgments, therefore improving judgment reliability and accuracy, by providing examples of observable behaviors along different points of a rating scale; a separate BARS is developed for each targeted job dimension (Bernardin & Smith, 1981). These behavioral anchors help to ensure that raters have a more standardized and uniform understanding of performance, which should result in more consistent and accurate interpretations and evaluations (Bernardin & Beatty, 1984; Schultz & Zedeck, 2011).

BARS were initially proposed by Smith and Kendall (1963) as a methodology to support more objective supervisory ratings of employee job performance than those produced using commonly used Likert scales that are anchored by adjectives. Likert scales are more prone to suffering from rater errors, such as leniency or halo