Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

Teacher evaluation: Are principals' classroom observations accurate at the conclusion of training?

Christi Bergin[a,*], Stefanie A. Wind[b], Sara Grajeda[c], Chia-Lin Tsai[d]

[a] University of Missouri, 323 Clark Hall, Columbia, MO 65211, United States
[b] The University of Alabama, Box 870231, Tuscaloosa, AL 35487, United States
[c] University of Delaware, 201C Willard Hall, Newark, DE 19711, United States
[d] University of Missouri, 2800 Maguire Blvd, Columbia, MO 65211, United States

ARTICLE INFO

Keywords:
Teacher evaluation
Evaluators
Evaluation methods
Classroom observation
Observation of teaching practice
Rasch models

ABSTRACT

Teacher evaluation commonly includes classroom observations conducted by principals. Despite widespread use, little is known about the quality of principal ratings. We investigated 1,324 principals' rating accuracy of six teaching practices at the conclusion of training within an authentic teacher evaluation system. Data are from a video-based exam of four 10-minute classroom observations. Many-Facet Rasch modeling revealed that (1) overall principals had high accuracy, but individuals varied substantially, and (2) some teaching episodes and practices were easier to rate accurately. For example, promotes critical thinking was rated more accurately than uses formative assessment. Because Many-Facet Rasch modeling estimates individuals' accuracy patterns across teaching episodes and practices, it is a useful tool for identifying areas that individual principals, or groups, may need additional training (e.g., evaluating formative assessment). Implications for improving training of principals to conduct classroom observations for teacher evaluation are discussed.

A common approach to evaluating teachers' effectiveness is classroom observation of teaching practice (OTP) by supervising principals (Goe, Bell, & Little, 2008; Herlihy et al., 2014). These observations, and by extension, teacher evaluations, serve at least two purposes: to contribute to high-stakes, summative teacher evaluations, and to provide formative feedback to improve teaching. Yet, there is still relatively little empirical evidence to support the use of OTP ratings for either purpose, especially in authentic contexts, despite the high stakes associated with them.

This study seeks to answer the call for more research in this area (e.g., Cohen & Goldhaber, 2016) by investigating the accuracy of principals at the conclusion of OTP training within an authentic evaluation system. Accurate OTP ratings reflect a teacher's true effectiveness rather than idiosyncrasies in principal judgments (e.g., biases and other rating errors) and lack of training or expertise applying the observation protocol. Inaccurate ratings are unfair to teachers, and provide misinformation on teachers' effectiveness globally as well as misidentify particular strengths and areas needing growth, thereby failing both purposes of teacher evaluation. Inaccurate ratings are ethically unacceptable for high-stakes personnel decisions (AERA, APA, & NCME, 2014). Recently, a team of psychometricians argued that we need to ensure that "ratings assigned by raters [such as principals] are accurate, consistent with scoring protocols, and free of bias. . to appropriately assess teacher performance" (Sukin et al., 2014).

Ideally, we need to ensure that ratings in the field, not just at the conclusion of training, are accurate. However, few, if any, authentic evaluation systems have the resources to investigate the accuracy of in-field ratings where typically a single principal evaluates many teachers and no two principals evaluate the same teacher. Investigation of the accuracy of ratings at the conclusion of rater training is an important first step because accuracy at this point is foundational to accuracy in the field.

This study also seeks to answer the call for more research in this area by demonstrating an approach to assessing the accuracy of OTP ratings that provides diagnostic information about individual principals, teaching episodes, and teaching practices. Such information is critical in order to inform the interpretation and use of OTP ratings, as well as to improve practice in training principals for OTP. Our approach has implications for analyzing the training process and for raising concerns relevant to in-field ratings, such as identifying whether some teaching practices are harder to rate accurately. It can be applied across teacher evaluation systems.

This study uses a criterion-referenced approach to evaluating principal accuracy in OTP. Different approaches have been developed

* Corresponding author.
E-mail addresses: berginc@missouri.edu (C. Bergin), swind@ua.edu (S.A. Wind), sbchap@udel.edu (S. Grajeda), tsaic@missouri.edu (C.-L. Tsai).

http://dx.doi.org/10.1016/j.stueduc.2017.05.002
Received 14 December 2016; Received in revised form 13 April 2017; Accepted 16 May 2017

to assess rater accuracy that reflect varying definitions of accuracy for performance assessments. For example, in Generalizability theory, high reliability coefficients—indicating consistency of teacher rankings across raters—are considered evidence of rating accuracy (Brennan, 2000). Other approaches compare ratings of operational raters against criterion ratings of experts who have extensive experience with the assessment system, such that alignment between operational and criterion ratings are considered evidence of rating accuracy. In the few studies in which the quality of OTP ratings have been assessed, they typically use reliability coefficients (e.g., Ho & Kane, 2013; Kane & Staiger, 2012). Reliability coefficients are difficult to interpret regarding the quality of rater judgments. For example, large coefficients suggest that principals provide consistent rankings of teachers, yet consistency does not necessarily imply accuracy. Furthermore, while reliability is potentially appropriate in contexts focused on relative standing, investigating rating accuracy from a criterion-referenced perspective is more appropriate in contexts where scores have specific meaning (e.g. earning a score of “5” identifies teachers as “highly effective”).

Several scholars have incorporated a criterion-referenced approach into modern measurement techniques based on latent trait models (i.e., item response theory models). For example, Engelhard (1996), Wind and Engelhard (2013), and Wolfe, Song, and Jiao (2016) showed how Many-Facet Rasch (MFR) models (Linacre, 1989) can be used to systematically evaluate rater accuracy based on the alignment between operational and criterion ratings. Specifically, rater accuracy, as defined by the match between operational and criterion ratings, is used as the dependent variable. Then, measures of rater accuracy and the difficulty associated with accurate ratings for examinee performances and other facets can be estimated. These accuracy estimates reflect the overall scoring accuracy of individual raters, and the difficulty associated with providing accurate ratings on particular facets, such as teaching practice or teaching episode. Other facets can be included in the model in order to examine the difficulty of assigning accurate ratings related to additional aspects of an assessment system, such as rubric domains. Previously the MFR approach has primarily been used to evaluate rating quality for writing performance assessments. This study extends the use of MFR modeling to a teacher evaluation context to inform the improvement of measures, rater training practices, and other components of teacher evaluation systems.

This study addresses three research questions in the context of training principals for accuracy in an authentic evaluation system: (1) How accurate are principals at the conclusion of training, and does rating accuracy vary across principals? (2) Does rating accuracy vary by teaching episode or teaching practice? (3) Does the MFR model yield helpful diagnostics to inform training within teacher evaluation systems?

1. Methods

1.1. Participants

This study explores data from principal training for OTP in summer of 2015. Principals had between one and five years of experience conducting OTP in their own schools. All principals (n = 1324) who completed the exam were included in the data. Participants were 50.3% female. Principals of elementary schools (39.6%), secondary schools (40.5%), both elementary and secondary schools (9.9%), and alternative or early childhood centers (10.0%) were included. Participating principals represented schools from urban to rural and high- to very low-income students. Thus, the principals lead a diverse cross-section of schools.

1.2. Setting and training procedure

This study draws upon a rich state-wide database. Data were collected through the Network for Educator Effectiveness (NEE), which is a teacher evaluation system used by over 265 diverse school districts across the state of Missouri. NEE was developed in collaboration between practitioners and researchers at the University of Missouri.

Principals participate in annual teacher evaluation trainings in groups of 20 to 30 during each summer. Training is carefully designed to follow best practices. NEE uses a “rater error” training approach in which raters are trained to recognize and avoid making leniency errors and halo errors, and to use the full scale. Raters are trained to begin with a middling rating of “3” and then only move up or down the scale if the evidence clearly justifies doing so. NEE also uses a “performance dimension” training approach in which raters learn to understand common teaching practices through discussion and literature review. Finally, NEE also uses a “practice-with-feedback” training approach which asks raters to watch and rate carefully selected videos of authentic classes that portray a range of ratings (across a range of subjects and grade levels). They first view and rate videos on their own, then justify their ratings in small groups, and then share with a large group. Trainers give additional feedback based on criterion ratings of the practice videos. Together these training approaches should reduce error and increase accuracy (Chafouleas, 2011; Woehr & Huffcutt, 1994). Principals then take a video-based exam at the conclusion of training. As members of the NEE network, principals are expected to conduct 10-min, unannounced OTP ratings 6–10 times per school year of every teacher in their buildings.

1.3. Measure

The NEE classroom observation rubric is based on the Interstate Teacher Assessment and Support Consortium (InTASC) standards (Council of Chief State School Officers, 2011), as condensed by the Missouri State Department of Elementary and Secondary Education. Principals assign a rating from 0 (not present) to 7 (perfect exemplar) for each teaching practice. On the NEE rubric, anchor ratings (i.e. 0, 1, 3, 5, and 7) have clear, specific behavioral descriptions. Ratings are given for each teaching practice separately, so a teacher may be assigned a rating of “2” on “promotes critical thinking” but a “6” on “uses formative assessment.”

Four 10-min videos of authentic classrooms were included in the exam. Each video depicted a different teaching episode: (1) 5th-grade language arts, (2) 4th-grade math, (3) High School International Baccalaureate (IB), and (4) 9th-grade math. These videos were selected to reflect a range of grade levels, subject areas, and teaching effectiveness. Principals completed the exam at a personal computer station at the training site, using headphones. Principals rated the teachers in each episode on six teaching practices: (1) Use of academic language, (2) Cognitive engagement, (3) Critical thinking, (4) Motivation, (5) Teacher-student relationships, and (6) Formative assessment. Principals took notes on paper forms at their station, and then recorded their rating into a Qualtrics survey.

Principals’ ratings were compared to criterion ratings that had been established by the rubric developers and a selected group of “expert raters”; principals who had experience scoring at least 75 OTPs for each teaching practice in their buildings. To obtain criterion ratings, between three and six expert raters watched and rated the videos independently, followed by a small group discussion to justify scores and resolve discrepancies. Criterion ratings were established based on the results of two groups of expert raters to ensure scores were robust. Principals were considered accurate if they had adjacent agreement (within plus or minus one) with the criterion rating on the 8-point scale.

1.4. Data analysis

This study uses a Many-Facet Rasch (MFR) model to explore OTP scoring accuracy based on a match between operational and criterion ratings. First, principal ratings on the qualifying exam were classified as