Studies in Educational Evaluation 55 (2017) 19–26

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

either accurate (*Accuracy rating* = 1) if they were within one point of each criterion rating, or inaccurate (*Accuracy rating* = 0) otherwise. This dichotomous classification reflects previous applications of the MFR model to evaluate rater accuracy in performance assessments (Engelhard, 1996; Wind & Engelhard, 2013; Wolfe et al., 2016). It also reflects the approach used in the NEE system during training and for qualifying principals to use the system.

The model used in this study was adapted from the MFR model for rater accuracy proposed by Engelhard (1996). Rather than modeling observed ratings, this model uses the match between principal ratings and criterion ratings as a dichotomous dependent variable that reflects principal accuracy. The model can be specified to reflect a variety of facets in the exploration of scoring accuracy. Analyses were conducted using the Facets computer program (Linacre, 2015).

The model used here included three facets: principals (n = 1324), teaching episodes (n = 4), and teaching practices (n = 6). Stated mathematically, the model is:

$$\ln\left[\frac{P_{nij(x=1)}}{P_{nij(x=0)}}\right] = \beta_n - \lambda_i - \delta_j, \quad (1)$$

where:

$\ln[P_{nij(x=1)}/P_{nij(x=0)}]$ = the probability that Principal n provides an accurate rating of Teaching Episode i related to Teaching Practice j, rather than an inaccurate rating,

β_n = the accuracy level of Principal n,

λ_i = the difficulty of assigning an accurate rating to Teaching Episode i, and

δ_j = the difficulty of assigning an accurate rating to Teaching Practice j.

When Eq. (1) is used to model the dichotomus accuracy ratings, a variety of statistics and displays are obtained that describe the calibration of the principals, teaching episodes, and teaching practices on a common scale that represents the construct of principal scoring accuracy. Of particular interest in this study are the calibrations of the individual elements within each facet (individual principals, teaching episodes, and teaching practices). Model-data fit statistics are also calculated in order to verify the degree to which these calibrations and separation statistics can be interpreted.

1.4.1. Calibrations

First, the MFR model for scoring accuracy (Eq. (1)) provides estimates for each individual principal, teaching episode, and teaching practice on a common linear scale that represents the construct of principal scoring accuracy (the logit scale). For each individual element within each facet, a value on the logit scale is estimated that reflects the log of the odds for an accurate rating associated with that particular principal, teaching episode, or teaching practice.

1.4.2. Variable map

When data fit the model (described below), the logit-scale locations of each facet can be compared to those of the other facets in order to identify differences in terms of scoring accuracy. It is common practice in Rasch measurement theory to construct a visual display called a *variable map* that illustrates the calibrations of each facet on the logit scale. The variable map provides a useful summary of the overall results.

1.4.3. Separation statistics

Using results from Eq. (1), separation statistics can be calculated that describe the degree to which differences among individuals and items are observed in a measurement procedure. First, the *reliability of separation* statistic describes the degree to which individual elements within a facet can be differentiated from one another, such as individual principals, teaching episodes, and teaching practices. For the object of measurement (in this case, principals), the reliability of separation is comparable to Cronbach's alpha coefficient when data fit the model, because it reflects an estimate of true-rating to observed-rating variance. For the other facets, the reliability of separation statistic describes the spread, or differences in the difficulty associated with providing an accurate rating across teaching episodes and teaching practices. Second, a *chi-square statistic* (χ^2) is calculated that describes the degree to which the logit differences within each facet (individual principals, teaching episodes, and teaching practices) are statistically significant.

1.4.4. Model-data fit

Model-data fit statistics are used within Rasch measurement theory to examine the degree to which adherence to the requirements for invariant measurement is observed in a set of data. In this study, model-data fit statistics are examined for the three facets in order to support the interpretation of calibrations of each facet as indicators of scoring accuracy. A variety of fit statistics have been proposed for use with Rasch models (Smith, 2004). This study uses two statistics calculated in the Facets computer program: Infit and Outfit mean square error (*MSE*), and standardized versions of the Infit and Outfit statistics. In this study, unstandardized (*MSE*) and standardized fit statistics are calculated for each principal, teaching episode, and teaching practice.

1.4.5. Accuracy profile plots

In addition to calibrations, separation statistics, and model-data fit indices, this study uses visual displays of observed principal accuracy as an additional diagnostic tool for exploring patterns in accuracy across teaching episodes and teaching practices for individual principals.

2. Results

First, summary statistics are presented that describe the overall MFR model calibration of the principal, teaching episode, and teaching practice facets, along with separation statistics for each facet. Next, the variable map is presented, and differences in principal scoring accuracy related to the teaching episodes and teaching practices are discussed. Finally, a few accuracy profile plots are presented in order to illustrate accuracy patterns for individual principals.

2.1. Summary statistics

Table 1 presents summary statistics that describe the overall results from the MFR model for principal accuracy including average calibrations, model data fit statistics, and reliability of separation statistics. In order to facilitate the interpretation of values on the logit scale, the teaching episode and teaching practice facets are centered (M = 0), and the principal facet was allowed to vary. As Table 1 shows, the average principal accuracy calibration was 1.52 (SD = 1.04), which is higher than the average calibrations of teaching episodes (M = 0.00, SD = 0.54) and teaching practices (M = 0.00, SD = 0.56). This result suggests that participating principals had high overall accuracy.

Model-data fit statistics for each facet indicate adequate fit to the MFR model. This finding supports the interpretation of the facet calibrations on a common linear scale that represents the construct of scoring accuracy.

Finally, separation statistics indicate differences in logit scale locations for the individual principals, teaching episodes, and teaching practices. The reliability of separation statistic for the principal facet was slightly lower (*Rel* = 0.61) than the value observed for the teaching episode and teaching practices facets (*Rel* = 0.99). When examined alongside the significant values of the chi-square statistic for these three facets ($p < 0.001$), these findings suggest that there are significant differences between individual principals, teaching episodes, and teaching practices. The finding of a slightly lower reliability of separation statistic for the principal facet suggests that there are clusters of principals with similar accuracy levels that significantly

Table 1
Many-Facet Rasch Model Summary Statistics.

	Principals	Teaching Episodes	Teaching Practices
Measures			
M	1.52	0.00	0.00
SD	1.04	0.54	0.56
N	1,324	4	6
Infit MSE			
M	1.01	1.00	1.00
SD	0.15	0.10	0.09
Std. Infit			
M	0.10	−0.10	−0.52
SD	0.70	0.92	1.27
Outfit MSE			
M	0.95	0.95	0.95
SD	0.29	0.19	0.18
Std. Outfit			
M	0.00	−0.38	−0.78
SD	0.70	1.00	1.86
Separation Statistics			
Reliability of Separation	0.61	0.99	0.99
χ^2 Statistic	3,296.8*	1,110.4*	1,743.3*
Degrees of Freedom	1,323	3	5

* $p < 0.001$.

differ from the accuracy of other principal clusters.

2.2. Variable map

2.2.1. Principal variability

Fig. 1 is the variable map that corresponds to the summary statistics presented in Table 1. The first column is the logit scale, which serves as the operational definition of the principal scoring accuracy construct. Higher logit scale values correspond to higher accuracy—captured through frequent matches between principals' and criterion ratings. On the other hand, lower logit scale values correspond to lower accuracy. The next column displays principal locations on the construct, with principals ordered from bottom to top in terms of increasing accuracy levels. Principals with low logit-scale locations tend to provide inaccurate ratings, and principals located higher on the logit scale tend to provide accurate ratings. Each star (*) represents 19 principals, and each period (.) represents between one and 18 principals. Overall, these principal calibrations indicate a wide spread of accuracy across the 1324 participating principals, with accuracy measures ranging from −1.50 logits for the least accurate principal (*Observed average accuracy rating* = 0.22), to 4.68 logits for the most accurate principal (*Observed average accuracy rating* = 0.99).

2.2.2. Teaching episode variability

The next column displays the locations of the four teaching episodes on the logit scale. For this facet, low measures on the logit scale indicate that a teaching episode is easy to rate accurately, and high measures on the logit scale indicate that a teaching episode is difficult to rate accurately. Starting at the bottom of the variable map, Grade 4 Math was the easiest teaching episode to rate accurately (*Measure* = −0.83 logits, *Observed average accuracy rating* = 0.87), followed by Grade 9 Math (*Measure* = −0.10 logits, *Observed average accuracy rating* = 0.78), High School IB (*Measure* = 0.35 logits, *Observed average accuracy rating* = 0.72), and Grade 5 Language Arts (*Measure* = 0.58 logits, *Observed average accuracy rating* = 0.68).

Table 2 (Panel A) shows differences in the logit scale locations for these teaching episodes. The chi-square statistic for the teaching episode facet indicates significant overall differences across the four teaching episodes ($p < 0.001$). Following Engelhard and Myford (2003), differences between logit calibrations that exceed |0.30| are interpreted as substantively meaningful. Such differences in scoring accuracy occurred between all pairs of teaching episodes except between Grade 5 Language Arts and High School IB (*Difference* = 0.23 logits).

2.2.3. Teaching practice variability

Similarly, the fourth column shows calibrations of the six teaching practices on the logit scale. Again, low measures on the logit scale indicate that a teaching practice is easy to rate accurately. Starting at the bottom of the variable map, the easiest teaching practice to rate accurately was 4.1: *The teacher uses instructional strategies to get students to problem solve and think critically* (*Measure* = −0.53 logits, *Observed average accuracy rating* = 0.84), followed by teaching practice 5.1: *The teacher uses motivation strategies effectively* (*Measure* = −0.30 logits, *Observed average accuracy rating* = 0.81). The most difficult teaching practice to rate accurately was 7.4: *The teacher conducts on-going assessment of learner progress during the lesson* (*Measure* = 1.21 logits, *Observed average accuracy rating* = 0.56). Only small differences in logit-scale locations were detected for the remaining three teaching practices. Among these, teaching practice 1.1: *The teacher uses academic language to communicate key concepts of the discipline and gets students to use academic language accurately* was easiest to rate accurately (*Measure* = −0.15 logits, *Observed average accuracy rating* = 0.79), followed by teaching practice 1.2: *The teacher uses strategies to cognitively engage students* (*Measure* = −0.12 logits, *Observed average accuracy rating* = 0.79), and teaching practice 5.3b: *The teacher has positive interactions or has a positive relationship with students* (*Measure* = −0.11 logits, *Observed average accuracy rating* = 0.79).

Table 2 (Panel B) shows differences in the logit scale locations for these teaching practices. The chi-square statistic for the teaching practice facet indicates significant overall differences across the four teaching practices ($p < 0.001$). Substantively meaningful differences in scoring accuracy occurred for all comparisons involving teaching practice 7.4, which was the most difficult teaching practice to rate accurately. Other substantively meaningful differences involved teaching practice 4.1, which was the easiest teaching practice to rate accurately; these comparisons included teaching practice 1.1, 1.2, and 5.3b.

2.3. Accuracy profile plots

Plots can be constructed to illustrate accuracy patterns for individual principals. Principals who are accurate across teaching episodes or teaching practices would have consistently high average accuracy ratings. In contrast, principals whose accuracy varies across teaching episodes or teaching practices would have different accuracy averages across the levels of these facets.

Fig. 2 (Panel A) provides an illustrative plot for three principals (A, B, and C) that demonstrates differences in accuracy across teaching episodes. Profiles for the three principals are summarized across the six teaching practices within each teaching episode. The average accuracy rating is plotted along the *y*-axis, and the four teaching episodes are shown along the *x*-axis. Principal A has high average accuracy across all four teaching episodes based on the MFR model ($\beta = 4.68$). Principal B has an average overall accuracy based on the MFR model ($\beta = 0.50$) indicating varied accuracy across the four teaching episodes. This principal has greater accuracy when scoring the Grade 4 Math and High School IB episodes than when scoring the Grade 5 Language Arts and Grade 9 Math episodes. Finally, Principal C has a low overall accuracy based on the MFR model ($\beta = 0.21$) indicating varied accuracy across teaching episodes. This principal has greater accuracy when scoring the High School IB episode, compared to the other three episodes.

Similar accuracy profiles are presented in Fig. 2 (Panel B) for the six teaching practices. The same three principals are used to illustrate patterns of accuracy, and the interpretation of the profile plots is similar

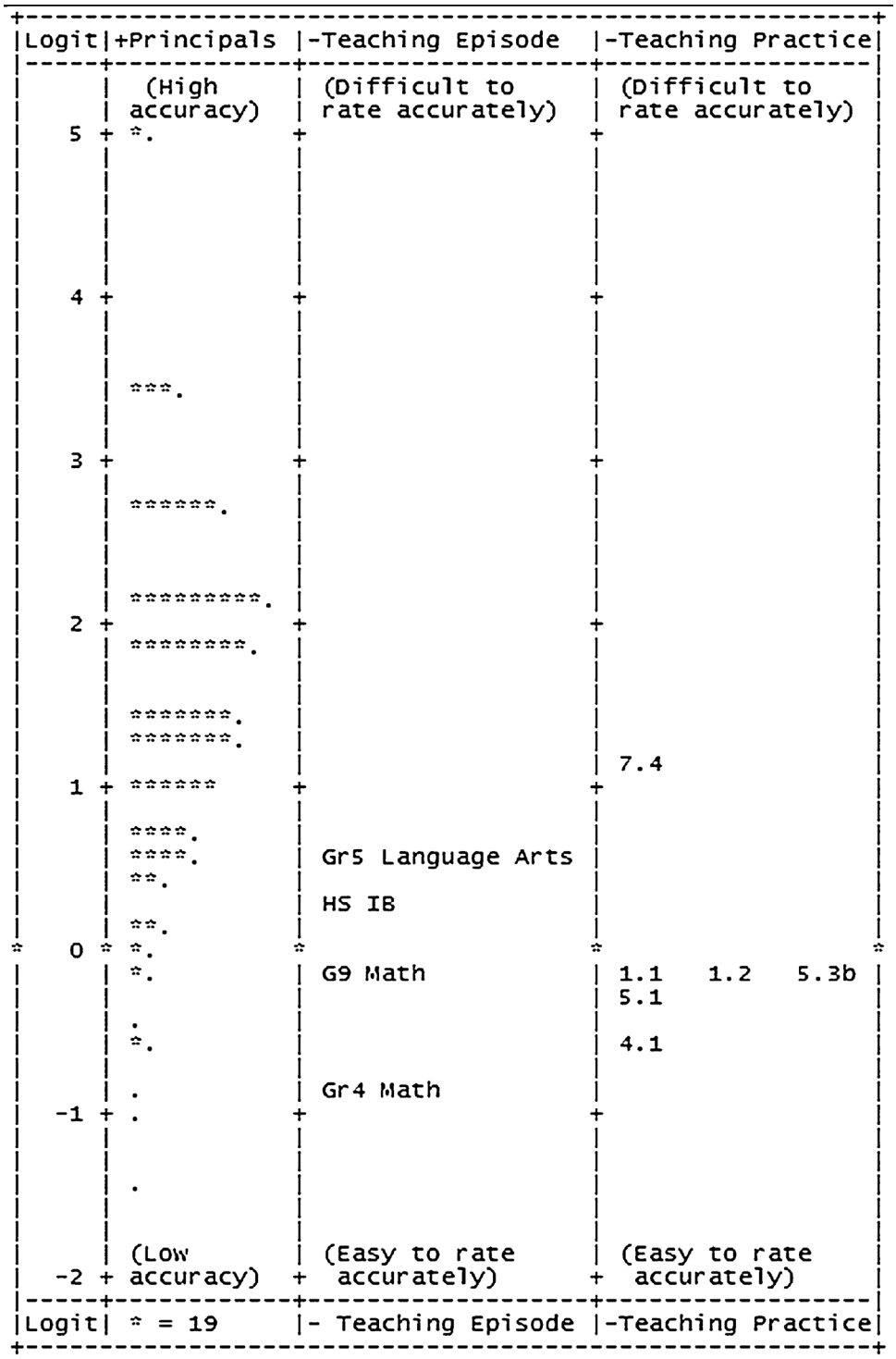

Fig. 1. Variable Map for Principal Accuracy. *Note*. In the principals column, a star (*) represents 19 principals, and a period (.) represents between one and 18 principals. Teaching practices are numbered according to the Network for Educator Effectiveness system: 1.1 = The teacher uses academic language to communicate key concepts of the discipline and gets students to use academic language accurately; 1.2 = The teacher uses strategies to cognitively engage students; 4.1 = The teacher uses instructional strategies to get students to problem solve and think critically; 5.1 = The teacher uses motivation strategies effectively; 5.3b = The teacher has positive interactions or has a positive relationship with students; 7.4 = The teacher conducts on-going assessment of learner progress during the lesson.

to the plot in Panel A. Principal A has consistently high accuracy across all six teaching practices, and Principals B and C have varied accuracy across the teaching practices.

3. Discussion

This study explored the accuracy of principals' OTP ratings at the conclusion of training in an authentic teacher evaluation system. This is a topic of national importance because many districts mandate OTP as part of teacher evaluation, yet little is known about the accuracy of such data. While ideally we want to know about principals' accuracy during in-field observations, investigation of their accuracy at the conclusion of training is an important first step. This study also explored the use of Rasch measurement theory as an innovative tool in teacher evaluation

Table 2
Differences in Principal Rating Accuracy.

A. Differences Related to Teaching Episodes

Teaching Episode	Measure	Mean Differences in Accuracy			
		Grade 4 Math	Grade 5 Language Arts	Grade 9 Math	High School IB
Grade 4 Math	−0.83	–	−1.41*	-0.73*	-1.18*
Grade 5 Language Arts	0.58		–	0.68*	0.23
Grade 9 Math	−0.10			–	-0.45*
High School IB	0.35				–
χ^2	1,110.4*				
Degrees of Freedom	3				

B. Differences Related to Teaching Practices

Teaching Practice	Measure	Mean Differences in Accuracy					
		1.1	1.2	4.1	5.1	5.3b	7.4
1.1: Academic Language	−0.15	–	−0.03	0.38*	0.15	−0.04	−1.36*
1.2: Cognitive Engagement	−0.12		–	0.41*	0.18	−0.01	−1.33*
4.1: Critical Thinking	−0.53			–	−0.23	−0.42*	−1.74*
5.1: Motivation	−0.30				–	−0.19	−1.51*
5.3b: Relationships	−0.11					–	−1.32*
7.4: Formative Assessment	1.21						–
χ^2	1,743.3*						
Degrees of Freedom	5						

* $p < 0.001$.

for exploring rating accuracy that provides diagnostics on individual principals, different teaching episodes, and varied teaching practices. Key findings will be discussed next.

3.1. How accurate are principals' OTP ratings?

Principals had high overall accuracy, as indicated by the spread of principal calibrations on the logit scale and separation statistics for the principal facet. These results are promising in light of findings that "raters are the largest source of error" in OTP in the context of research studies with no consequences for the teachers being rated, such as the well-known MET study, as opposed to authentic evaluation systems (Cohen & Goldhaberg, 2016, p. 382). Our promising results may follow from using an innovative approach, rather than conventional reliability estimates, that models principal accuracy as a linear continuum on which individual principals are calibrated while accounting for additional facets (teaching practices and teaching episodes) that influence principal accuracy. Our promising results may also follow from key attributes of the NEE evaluation system, such as (1) principals receive three days of initial training and then one day of re-calibration training annually, (2) the rubric is quantitative (not categorical) with an elongated 8-point scale that has behavioral descriptors, and (3) there is no "cut" score for classification as proficient (the presence of cut scores tends to result in teachers stacking up just over the cut score). Further research is needed to determine which attributes of teacher evaluation systems (e.g., the rubric, the training format or length) most affect OTP accuracy.

Principals varied significantly in accuracy; however, they tended to cluster. That is, a moderate reliability-of-separation statistic indicated that there are several groups of principals with similar accuracy, within which individual differences are not distinct. This means that while principals overall tend to be accurate, some teachers are being evaluated by principals who demonstrated low accuracy, and some are being evaluated by principals who demonstrated high accuracy at the conclusion of training. Teachers could be advantaged or disadvantaged by inaccurate ratings depending on whether a principal's inaccuracy results in artificially high or low ratings in the field. Further research in which accuracy is defined using a polytomous rating scale or an unfolding approach will provide additional insight into the directionality of principal scoring accuracy.

Attributes of principals may influence accuracy of OTP, such as teaching experience in particular subject areas and grade levels, along with the match between principal characteristics and characteristics of teaching episodes. For example, new principals may not be as accurate as principals who have conducted OTP for teacher evaluation for years. (In the present study, all principals had been in the NEE system for at least a year.) For another example, a principal who is a former elementary math teacher may more accurately score an elementary math lesson than a high school social studies lesson. Additional research should address the effect of these principal attributes on OTP accuracy. Nevertheless, current practice in teacher evaluation often requires principals to evaluate teachers across a wide range of subject and grade levels, such that this study reflects the reality of evaluation systems. For example, one NEE principal of a rural K-8 school is a former high school coach, but now has to evaluate primary teachers.

3.2. Does rating accuracy vary by teaching episode or teaching practice?

Teaching episodes varied in how difficult they were to rate accurately. The Grade 4 math episode was significantly easier to rate accurately than the Grade 9 math episode, which was easier to rate than the High School IB or Grade 5 Language Arts episodes. The latter two episodes did not differ in difficulty. Because the four episodes varied by both subject and grade, we cannot determine whether these variables or other lesson-specific variables might affect accuracy. However, this finding has important implications for teacher evaluation because it suggests that lessons may vary in rating difficulty. Future research should address what variables affect OTP accuracy. Variation due to subject or grade suggests that teacher evaluation data should only be compared within similar subjects and grades. Variation due to lesson idiosyncrasies suggests that teacher evaluation should be based on multiple lessons rather than a single lesson. Future research should also address teaching episodes across grade levels and subject areas that were not included in the current study.

Teaching practices varied in how difficult they were to rate accurately although each teaching practice was measured within the same teaching episode and rubric scale. Pairwise comparisons revealed significant differences between the most difficult-to-rate teaching practice (7.4: *Formative Assessment*) and the easiest-to-rate teaching practice (4.1: *Critical Thinking*). These results support NEE trainers' and participant principals' anecdotal experience that principals struggle with identifying the effectiveness of formative assessment. In response, this teaching practice will be emphasized during training in the coming year. Subsequent analysis will indicate whether this emphasis results in greater accuracy. Interestingly, in previous years there was a similar struggle among principal trainees about how to identify the promotion of critical thinking. In response, NEE trainers focused on what critical thinking is (e.g., a reasoned argument or solving an ill-structured problem) and is not (e.g., spouting opinion or routine use of an algorithm), and now critical thinking is the easiest-to-rate teaching practice. This suggests that, although some teaching practices may be inherently challenging to rate accurately, emphasis and clarity in training may increase their accuracy.

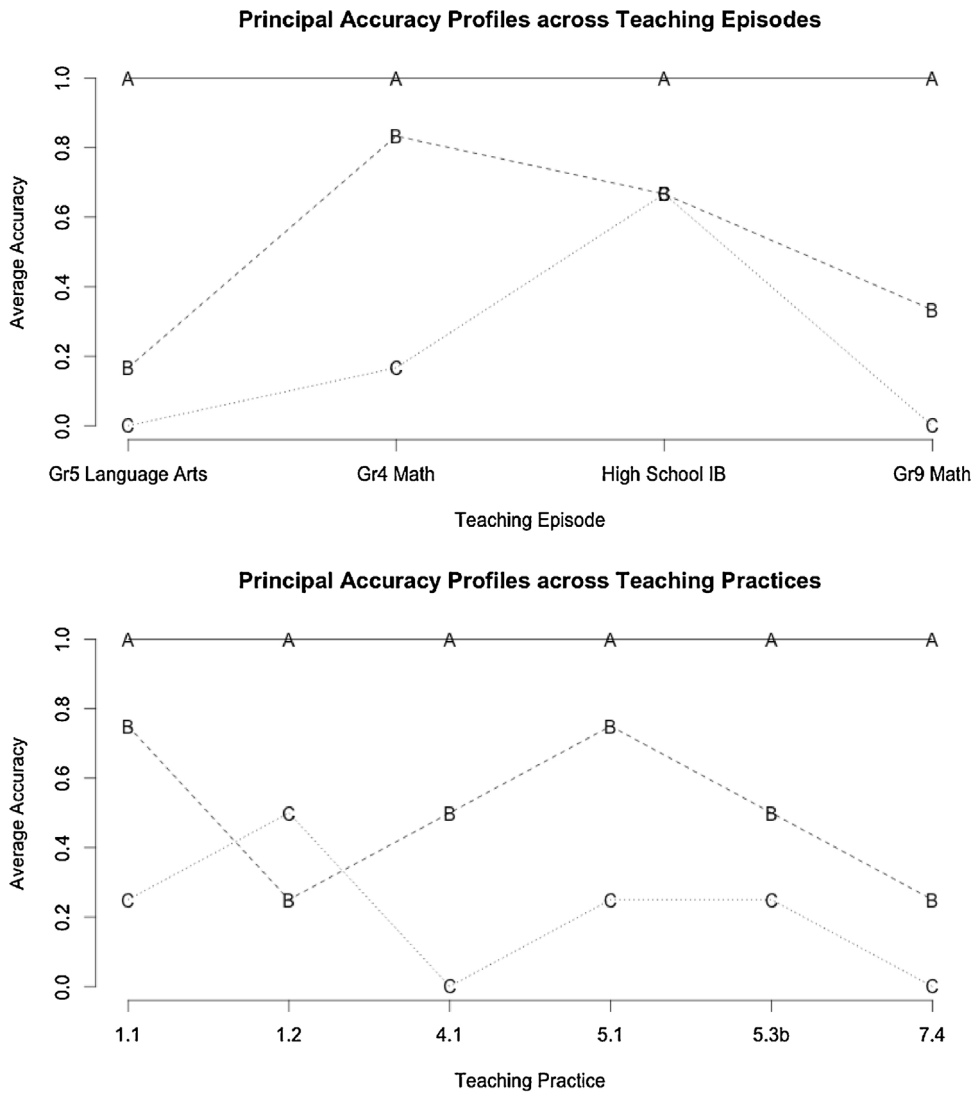

Fig. 2. Principal Accuracy Profiles. *Note.* Lines (A, B, and C) represent three different principals. Teaching practices are numbered according to the Network for Educator Effectiveness system: 1.1 = The teacher uses academic language to communicate key concepts of the discipline and gets students to use academic language accurately; 1.2 = The teacher uses strategies to cognitively engage students; 4.1 = The teacher uses instructional strategies to get students to problem solve and think critically; 5.1 = The teacher uses motivation strategies effectively; 5.3b = The teacher has positive interactions or has a positive relationship with students; 7.4 = The teacher conducts on-going assessment of learner progress during the lesson.

3.3. Does the Many-Facet Rasch model yield helpful diagnostics to inform training?

The MFR model reveals which principals rate more or less accurately. Furthermore, *accuracy profile plots identify which teaching episodes and teaching practices each individual principal finds more or less difficult to rate accurately*. This diagnostic tool could inform professional development or coaching customized to individual principal needs. For example, some principals may need additional training in a specific teaching practice (e.g., promoting critical thinking) or a specific subject area (e.g., guided reading) that most others do not need.

The MFR model also reveals which teaching episodes and teaching practices are more or less difficult to rate accurately. This information could be useful in various ways. It could inform professional development across a broad cohort. For example, in the state of Missouri principals may need additional training in formative assessment. It could also be used to explore what variables make one teaching episode more difficult to rate accurately than another. Is it the grade level, subject (e.g., band vs. social studies), lesson format (e.g., whole group discussion vs. individual computer use), or other factors? Given that personnel decisions are based on OTP ratings, it is important to learn how to make them as accurate as possible. Future research should build on this study by investigating how to best increase accuracy of difficult-to-rate teaching practices and which teaching practices are consistently more difficult to rate accurately using a wider variety of teaching episodes and practices.

3.4. Limitations and directions for future research

This study has three important limitations that point to directions for future research beyond those already discussed above. First, the data were based on an exam at the end of training, using videos. Principals presumably took the exam seriously because they may not use the evaluation system if they do not perform well; thus, there was peer pressure to provide high-quality ratings. Yet, results of this study may not generalize to in-field OTP ratings where other concerns of principals (e.g., keeping teachers motivated, personal relationships) may take precedence over rating accuracy (Cohen & Goldhaber, 2016).

Second, accuracy scores were assigned dichotomously – accurate or inaccurate – based on adjacent agreement. Although this approach has been used in previous applications of the MFR model to examine rater accuracy, and this approach is used in the evaluation system in

which the data is embedded, it could result in a higher proportion of accurate ratings compared to an approach that bases accuracy on exact agreement. On the other hand, exact agreement may not be realistic on an 8-point rating scale, as used in this study. In teacher evaluation systems that use truncated scales, accuracy could be higher given that almost all teachers earn the same rating (Weisberg, Sexton, Mulhern & Keeling, 2009).

Finally, the NEE evaluation system is designed to inform individual or school-wide professional development. That is, the desired outcome is teacher effectiveness, not a rank-ordering of teachers based on competency. As a result, a criterion-referenced approach toward principal accuracy is more meaningful than a norm-referenced approach, such as reliability coefficients. Future research may compare criterion and norm-referenced approaches.

3.5. Conclusions and implications

In conclusion, within a large, authentic evaluation system that provides face-to-face training and a high-quality observation rubric, principals were overall accurate in their OTP ratings, relative to criterion ratings, at the conclusion of training. However, there was individual variation suggesting that some principals have low accuracy and others have high accuracy. If these results extend to in-field OTP, some teachers could be advantaged or disadvantaged by differences in rater accuracy related to idiosyncrasies in rater judgments, including biases and erroneous application of the scoring rubric. In addition, some teaching episodes and some teaching practices (e.g., use of formative assessment) were more difficult to rate accurately. These findings have implications for training principals to accuracy in teacher evaluation systems. MFR models can be used to identify specific areas in which individual principals may need additional training and to identify teaching practices that need increased focus during training.

Another key implication from this study is related to methodology. Although this was an innovative application of the MFR model, it proved to be a useful tool in the context of teacher evaluation because it goes beyond traditional indicators of reliability and provides diagnostic information to improve training based on a criterion-referenced perspective on accuracy. When the MFR model is used to explore rating accuracy, principal accuracy can be compared across facets such that accuracy of individual principals, teaching episodes, and teaching practices can be examined in relation to each of the other facets. The utility of the MFR model is not limited to the NEE system, but could be used in any teacher evaluation system in which a criterion-referenced accuracy indicator is available, whether that is expert ratings or average ratings across a sample of principals. In any such system the MFR model can be used to explore differences in principal accuracy related to various aspects of the evaluation system unique to that system.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Authors.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339–353. http://dx.doi.org/10.1177/01466210022031796.

Chafouleas, S. M. (2011). Direct Behavior Rating: A review of the issues and research in its development. *Education and Treatment of Children*, *34*(4), 575–591.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378–387. http://dx.doi.org/10.3102/0013189X16659442.

Council of Chief State School Officers (2011). *Interstate Teacher Assessment and Support Consortium InTASC Model Core Teaching Standards: A resource for state dialogue*. Washington, DC: Author.

Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*. New York: College Entrance Examination Board.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center of Teacher Quality.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., et al. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*(1), 1–28.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel. Research paper. MET project*. Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. Policy and practice brief*. Bill and Melinda Gates Foundation.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2015). *Facets Rasch measurement (Version 3.71.4)*. Chicago, IL: Winsteps.com.

Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith, & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Sukin, T., Nicewander, A., Winter, P., Mitzel, H., Keller, L., & Schulz, M. (2014). Take the time to evaluate teacher evaluation. *Education Week*, *33*, 28–29.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., et al. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: New Teacher Project.

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. http://dx.doi.org/10.1016/j.asw.2013.09.002.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1–10. http://dx.doi.org/10.1016/j.asw.2015.06.002.