Rethinking Rater Effects When Using Teacher Observation Protocols

Ying Chen, Rachel A. Yim, Richard Kogen, Mike Stieff and Alison Castro Superfine
ychen406@uic.edu, ryim@uic.edu, rkogen1@uic.edu, mstieff@uic.edu, amcastro@uic.edu
University of Illinois at Chicago

Abstract: Analysis of teacher practice using observation protocols is often adversely impacted by construct-irrelevant sources of variance. Identifying and understanding these factors and controlling their effects are crucial for improving the validity and reliability of observation protocols. Here, we describe the application of the many-faceted Rasch model (MFRM) to account for these effects in learning sciences research. We used data from 172 high school chemistry classroom videos to illustrate the utility of MFRM to model latent constructs related to raters. The analysis shows our raters had high internal consistency with varying levels of bias despite intensive training. Such differences can significantly impact claims about teacher practice. We demonstrate how MFRM can incorporate these differences into a model that accounts for rater biases and other threats to validity. Additionally, the model allows for the incorporation of individual raters' expertise into measures of teacher practice by accounting for identified variability without removing it.

Keywords: teacher practice, observation protocols, many-faceted Rasch model, validity, mixed methods

Introduction

Inquiry into teacher practice has been a longstanding component of work in the learning sciences. The community has abandoned design frameworks that attempt to "teacher-proof" classroom learning environments and shifted its focus to better account for teacher input into the design of learning environments and to better characterize the complexity of teacher practice (Gomez, Kyza & Mancevice, 2018). Necessarily, research in this area involves collecting large amounts of data on pedagogical practices employed in real time, typically through observations of classrooms. Borrowing from other disciplines, learning sciences researchers have increasingly turned to the use of observational protocols with rating scales that focus on a range of teacher practices and the nature of teacher-student interactions (Fishman, Davis & Chan, 2014). By design, such protocols reduce observations into a small number of quantitative variables that offer a simplified interpretation of a complex human activity. Of course, capturing the complexity of teaching is far beyond the capacity of one single statistical measure and is ultimately contingent on a variety of factors both within and outside classrooms; yet, the potential utility of observational protocols (e.g., data management, triangulation, multi-level modeling) merits their consideration in research on teacher practice. To that end, this paper argues for a novel technique for analyzing observational protocols that has the potential to improve rater training programs as well as account for rater variability in the analytical process.

Reliability and teacher observation protocols

Independent of questions around the validity of teacher observation protocols, investigators contend with a range of potential confounds to the scoring process. Among these confounds, issues related to the reliability of rater scoring is a central concern. The reliability of an implementation of a particular protocol is typically established through the use of inter-rater reliability metrics. After training, investigators assess correlations among rater scores to determine the reliability of their scoring procedures, most commonly via Cohen's kappa (Cohen, 1960). Using this method, investigators set a minimum threshold for reliability before interpreting aggregate scores. Scores that achieve this minimum are deemed reliable and those below are revisited, often in open dialogue, until raters reach an agreement. It is well established that such an approach may result in unintended interpretations of a scoring rubric (Eckes, 2008), biased ratings resultant from power dynamics in a research group (Hoyt & Kerns, 1999), or the need for costly and time-consuming training programs that often fail to produce a high degree of agreement (Barrett, 2001).

Beyond issues with rater training and agreement, this approach to inter-rater reliability (IRR) assumes that the underlying goal is to achieve an "objective" score for each teacher that is not subject to individual differences among raters (Eckes, 2011). We argue that there are two significant problems with this approach. First, this approach aims to negate the unique interpretation that an individual rater might have regarding specific

items on an observational protocol as well as differences in raters' expertise analyzing classroom environments. Second, this approach to IRR assumes that *rater drift,* which occurs when rater scores begin to vary over time, results not from important differences in the data, but from deviations in the rater's scoring due to their increased familiarity with a research participant or the observation protocol (Hoskens & Wilson, 2001). In practice, IRR and drift are often corrected by comparing an individual rater's score to implicit benchmarks or requiring raters to periodically recalibrate their interpretations in dialogue during the lifetime of a project (Myford & Wolfe, 2009).

In this paper, we argue that the prevailing techniques for establishing inter-rater reliability by homogenizing individual raters' scores may potentially occlude critical insights into teacher practice from individual raters. We argue that new approaches are necessary to establish the efficacy of rater training programs that also provide investigators with the flexibility to account for variation in rater scoring procedures. We present the application of the Many-Faceted Rasch Model (MFRM) as an alternative to using Cohen's kappa both to establish IRR and to account for rater variability and training. The technique offers researchers a way to develop systems and indices of teaching based on observational ratings of teacher practices (Johnson, Zheng, Crawford, & Moylan, 2019; Jones & Bergin, 2019). As such, it offers a unique contribution to investigations that can complement other data to present a fuller and more accurate characterization of teacher practice.

Many-faceted Rasch measurement models

For the past decade, psychometricians have been investigating the sources of rater variability using different types of techniques. Among these, Rasch models (especially MFRM) have been widely used, particularly on language assessments (McNamara & Knoch, 2012). Those studies have focused on the reliability of rater judgments, rater biases, and the relationship between rater bias and rater training. Though the majority of MFRM research has focused on high-stakes assessments, the applications have been extended to other research fields such as outpatient performance assessment (Kramer, Kielhofner, Lee, Ashpole & Castle, 2009), creative writing assessment (Barbot, Tan, Randi, Santa-Donato & Grigorenko, 2012), and behavior analysis (Mannarini, 2009). Here, we argue for its application to teacher observation protocols.

The many-faceted version of the Rasch measurement model is the extension of the ordinary 1PL (one-parameter) item response theory (IRT) model for polytomous items (Wright & Linacre, 1989). The basic mechanism of 1PL Rasch modeling is the transformation of ordinal observation data into linear logit measures. When both items and candidates are parameterized independently into a logistic regression, the probability of a participant's scores on an item is explained by the linear effects of the item difficulty and the relevant latent construct. When the item difficulty parameter is fixed, the item will demonstrate the same linear function to other participants with the same latent construct. With respect to teacher observation protocols, teacher practice is not solely determined by the item difficulties and the latent traits of participating teachers as mentioned in the 1PL Rasch model. As in any case of rater mediated data, threats to the reliability of the measurement result from rating scales, rating procedures, and the raters themselves (Myford & Wolfe, 2003). MFRM offers a mechanism to address these threats through modeling scaled responses that account for facets (e.g., rater bias, latent traits) that are known to have major group-level effects.

Similar to ANCOVA, MFRM includes within-group effects and rater effect interactions. Importantly, MFRM extends ANCOVA by separating out the facets. Thus, each facet can be analyzed individually without taking interaction effects into account. This allows for independently calibrating each facet on a logit scale. As a result, researchers using such a model can ultimately produce separate parameter estimations for teacher practice, rater bias, and item difficulty all on a single scale. Furthermore, the model provides an estimation of the contribution of each facet and calibrated weighted scores to determine whether facets function as predicted.

MFRM accomplishes this through an additive linear regression model with a common equal-interval metric (e.g., logit odd) transformation of observed scores (Wright & Linacre, 1989). It assumes that all items share a common rating scale classification, where the interval distance between each rating category functions identically across all items. Thus, the category thresholds do not vary across items. Equation (1) denotes a partial credit MFRM (Wright & Linacre, 1989) with four facets showing the log likelihood for teacher *n* to receive a rating of *k* in (*k-1*) stead on item *i* by rater *j* in classroom level *l* :

$$\ln\left[\frac{P_{nijlk}}{P_{nijlk-1}}\right] = \theta_n - \beta_i - \alpha_j - \gamma_l - \tau_{ik}, \quad (1)$$

Where

P_{nijlk} = probability of teacher *n* receiving rating *k* on item *i* in classroom level *l* by rater *j*,

$P_{nijlk-1}$ = probability of teacher *n* receiving rating *(k-1)* on item *i* in classroom level *l* by rater *j*

θ_n = level of performance for teacher *n*,
β_i = trait difficulty of item *i*,
α_j = bias (tendency to apply rating with more or less severity) of rater *j*
γ_l = traits of classroom *l* and difficulty of scale category *k* relative to scale category *(k-1)*.

Advantages of MFRM for analyzing teacher practice

MFRM offers two distinct advantages to researchers analyzing teacher practice over comparable techniques. First, MFRM produces a *reliability of separation* (strata) index. Using the separation statistics, it is possible to evaluate rater bias measures and identify whether there are differences among all raters across all items at both statistical and practical levels. First, this index could be used in lieu of Cronbach's alpha since one can examine observed score variance to identify raters who display inconsistent scoring practices or to differentiate raters from each other. Ideally this index should be located on a scale from 0 to 1 (Lumley & McNamara, 1995). Using the separation ratio investigators can calculate rater separation index (i.e., strata). This index categorizes raters into groups based on how similar they are in their rating practices (e.g., bias): When raters show a higher degree of variance, the strata will be greater. The strata are calculated by the separation index formula (Wright, 2002):

$$Strata = \frac{(4SD_t(H)+\sqrt{MSE_H})}{(3\sqrt{MSE_H})} = \frac{(4G_H+1)}{3}, \quad (2)$$

where $SD_t(H)$ is the observed variance of the rater's severity, MSE_H is the mean-squared measurement error, and G_H represents the rater separation ratio.

Second, MFRM produces discrete *rater fit statistics* (mean-squared statistics). In Rasch measurement, fit statistics present statistical evidence for how much variance in a dataset is expected based on the invariant measurement. In the rater facet, infit and outfit statistics are used to examine the consistency of raters. The unweighted fit mean-square (outfit) is calculated as the mean squared standardized residuals across all teachers and all items rated by each rater. Each rater's outfit statistics are sensitive to occasional unexpected ratings in a dataset with otherwise consistent ratings that are summarized over all facets:

$$MS_{outfit(j)=} \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nijl}^2}{NI}, \quad (3)$$

In contrast, each rater's weighted fit statistics (infit) provide evidence of an estimate of the consistency with which the rater uses the rating scale for all teachers across all items. Thus, the infit is more sensitive to outlying unexpected ratings that accumulate:

$$\mathrm{MS_{infit(j)=}} \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} w_{nij}^2 z_{nijl}^2}{\sum_{n=1}^{N}\sum_{i=1}^{I} w_{nijl}^2} \quad (4)$$

Where

$$w_{nij}^2 = \sum_{k=0}^{m}(k - e_{nij})^2 p_{nijlk}. \quad (5)$$

and z_{nijl}^2 represents the standardized score residuals.

Generally speaking, the infit statistic is more sensitive to an accumulation of unexpected ratings or model-data fit, while the outfit statistic is sensitive to individual unexpected ratings or outliers (Myford & Wolfe, 2003). Thus, outfit statistics can be used to evaluate the internal consistency of ratings to pinpoint outlying raters who might need additional training. Both infit and outfit estimations have an expected value of 1 and with the range of 0 to positive infinity (Wright, 2002; Wright & Linacre, 1989). Values greater than one indicate that a rater shows more variation than expected. Conversely, values less than one indicate a rater that shows less variation than model expects, which represents a rater who is likely to overfit the model. Higher values indicate the rater has more noise in their rating pattern than expected. Though there is no hard-and-fast rule to determine the lower bound and upper bound for the fit statistics, Wright and Linacre also suggest the infit and outfit value of 0.5 as a lower-control limit and 1.5 as an upper-control limit. In contrast, other researchers suggest using a narrower control limit from 0.7 to 1.3 when the instruments' interpretation is in a high stake test setting (Bond & Fox, 2001). Of course, the intervals in which the limits are defined must be considered in the context of the assessment purpose and available observations. In studies where inter-rater agreement is encouraged, previous

studies (e.g., Kaliski et al., 2013; Myford & Wolfe, 2004; Toffoli, de Andrade, & Bornia, 2016) suggest that mean-square values in the range between 0.4-1.2 produce "productive measurement".

Applying MFRM to teacher observation protocols: An illustration

We present here an application of the MFRM approach to a dataset from a longitudinal efficacy study of the Connected Chemistry Curriculum (CCC) (Stieff, Nighelli, Yip, Ryan, & Berry, 2012). CCC materials were designed to support teacher's adaptation of widely accepted, yet challenging, practices recommended by current reforms (National Research Council, 2012) that emphasize collaborative inquiry and scientific argumentation. Our goal was to analyze the reliability and identify sources of variability in the teacher observation protocol used in this study using MFRM. As we aimed to illustrate the application of MFRM and potential utility rather than evaluate the curriculum intervention, the rater training program, or instrument, we collected and analyzed the complied protocols from the project to identify variance in raters' scoring procedures.

Study context

Study participants included 20 high school chemistry teachers from three districts in a large urban area in the midwestern United States. Teachers in this study taught "on-level", "honors" and/or "AP" courses as defined by their districts. A team of six research assistants gathered and coded video recordings of each participant's chemistry lessons over a nine-month period using the Reformed Teaching Observation Protocol (RTOP) (Piburn et al., 2000). The RTOP was designed to measure the presence and extent of reform-based teaching practices at the lesson level. The version used was comprised of 26 five-point Likert scale items clustered into five categories (Lesson Design and Implementation, Propositional Knowledge, Procedural Knowledge, Communicative Interactions, and Student/Teacher Relationships). On the RTOP a rating of 0 indicates the absence of a practice described by each item, and a rating of 4 indicates that the description is highly present or characteristic of the lesson as a whole.

Procedure

The final dataset used in this analysis includes scored 172 RTOPs of 172 videos of classroom practice. Raters were trained to apply the RTOP rubric using 16 practice sessions and monthly calibration meetings where all raters discussed how to score a video as a group; the investigator goals for these calibration meetings was to achieve consensus (+/- 1 point) on a minimum of 80% of the RTOP items. In order to connect all the teachers in the analysis, at least two raters produced RTOPs for each teacher on the project using video collected from different lessons, and a single rater coded a minimum of one RTOP for all 20 teachers. Thus, the total number of scored items is 4456, with less than 0.01% (16) of missing items.

The FACETS program (v. 3.83, Linacre, 2019) was used to calibrate 4 facets in the dataset: teacher use of inquiry-oriented practices as defined by the RTOP, classroom level, rater bias, and item difficulty (see Table 1). The partial credit MFRM (see equation (1)) was used as a self-scaling structured evaluation where the threshold parameter is calibrated individually from item to item. All facets excluding the teacher facet (the target outcome variable of the RTOP) were centered to mean measure of zero and the joint maximum likelihood estimation (JMLE) for parameter estimation was used. The size of the maximum marginal residual was set at 0.5 logits. The JMLE estimation process converged after 252 iterations.

Results

Calibration of teachers, raters, items and classroom levels

Overall, there were 4456 valid responses used for parameters estimation in the partial-credit MFR model. 163 (3.65%) responses were outliers where responses were associated with standardized residuals equal or greater than 2 (absolute value), and ten responses (0.2%) were associated with standardized residuals equal or greater than 3 (absolute value). The model explains 61.75% of the variance in the RTOP scores. In summary, these results indicated an outstanding data-model fit overall. First, we present the results of the model before examining the utility of the model for examining how the model provides insight into rater bias, drift, and training.

The estimate of teachers' inquiry-oriented practices was found to vary from -1.01 logits up to 0.45 logits. With the mean measurement of teacher practice of -0.19 logit, most teachers (13 out of 20) were located in the -1.00 logit to 0.00 logit interval. Classroom level differences suggest a negligible influence on estimating a teacher's use of inquiry-oriented practices. "AP" classrooms were rated only marginally more difficult (0.04 logit) than "on-level" and "honors" classrooms (-0.02 logits). That is, teachers received similar estimates for using inquiry-oriented practices in all levels of classrooms.

The raters' bias facet captures how conservative raters' scoring practices were. The bias facet varied between -.0.63 logits to 0.49 logits with the mean calibrated to 0.00. Higher ratings on this scale indicate raters who are more severe in their rating practices and tend to provide lower scores on each RTOP item across teachers. For example, using the results in Table 2, one can expect that for any given teacher, Rater 6 will consistently apply a higher overall score on each RTOP item compared to Rater 4. As we could see from the output, the variance among raters was noticeable. With a spread of 1.12 logits, this facet accounts for 75% of the logit spread of variance in the model.

Lastly, the distribution of item difficulty ranged from -2.03 logits to 0.88 logits with a mean measure of 0. This range indicates two important findings regarding the instrument used in this study. First, the instrument covered the spectrum of teacher performance estimations, which indicates that the RTOP was a reliable instrument for analyzing teacher practices from a diverse group of teachers in this project. Second, the individual items on the instrument produce consistent and high-quality rating scales for measuring teachers' use of inquiry-oriented practices.

Table 1: Partial Credit MFRM Summary Statistics

	Teachers	Levels	Raters	Items
Descriptive Statistics				
N	20	3	6	26
Fit Measures				
Infit MNSQ	1.01(.23)	1.04(.08)	1.01(.14)	1(.09)
Outfit MNSQ	1.01(.25)	1.05(.11)	1.00(.14)	1(.11)
Other Measures				
Separation Ratio	4.73	0.00	10.54	7.55
Reliability	0.96	0.00	0.99	0.98
Strata	6.65	0.33	14.39	10.4
$\chi 2$	659.4(p=.00)	.6 (p=.75)	629.8(p=.00)	1296.5(p=.00)
Degrees of Freedom	19	2	5	25

Note. Standard deviation is presented in parentheses. Infit MNSQ= Infit Mean Square; outfit MNSQ= Outfit Mean Square

Insights into rater bias

Using the results of the model, we were able to analyze the rater group for bias in their coding procedures (e.g., the tendency of individual raters to provide higher or lower scores uniformly). Table 2 shows the distribution of rater bias. Among all distinctive raters in this study, Rater 4 stands out as being the most severe rater, having a biased measure of 0.49 logits; Rater 6 is the most lenient rater, having a biased measure of -0.63 logits. Thus, the scale suggests two important follow-up studies for investigators in this project. First, this pattern in the model could be produced from selection biases in the coding assignment: these raters may have coded a sample of teachers that were outliers among the entire group. Thus, the investigators should examine the coding process for evidence of asymmetric assignments. Second, this pattern might be produced from inadequate rater training. To examine this possibility, the investigators should review training procedures for these raters and employ benchmark coding assignments to ensure the two raters do not misunderstand the coding protocol.

Table 2: Rater Measurement Report

						MS		Average	
Rater	Obs%	Exp%	Discrim	Measure	S.E.	Infit	Outfit	Observed	Fair
Rater 1	29.90	33.70	0.82	-0.53	0.04	1.14	1.15	1.80	2.34
Rater 2	35.60	36.50	1.03	0.35	0.04	0.83	0.88	1.58	1.43
Rater 3	36.40	34.90	1.12	0.17	0.06	0.90	0.85	1.71	1.59

Rater 4	36.00	36.50	0.89	0.49	0.04	1.08	1.07	1.40	1.30
Rater 5	39.00	36.00	1.21	0.15	0.03	0.96	0.92	1.84	1.61
Rater 6	23.50	30.60	0.96	-0.63	0.05	1.17	1.14	2.03	2.45
Mean				0.00	0.04	1.01	1.00	1.73	1.79

Notes. Discrim= Discrimination estimates for raters.
Model, Population: RMSEA. = 04; Adj (True) S.D. = .42; Separation = 10.54; Strata = 14.39; Reliability (not inter-rater) = .99.
Model, Fixed (all same) chi-square: 629.8; d.f.: 5; significance (probability): .00.
Exact agreements: 2066, Obs% = 36.0%; Expected: 2051.3, Exp% = 35.7%.

Insights into rater drift

Table 2 presents rater fit statistics, which can be used to identify whether a specific rater produces unexpected ratings calibrated over all other facets. This index provides a mechanism to evaluate rater drift over the lifetime of the project. The infit and outfit statistics of all six raters are located in the suggested narrower 0.7 to 1.3 range and nested within relatively narrow intervals; this suggests there is a good fit to the model for the rater facet. According to the results, raters in this study were internally consistent and used the RTOP rating scale consistently over time. However, two raters (Rater 6 and Rater 1) exhibited noisy rating patterns. These two raters with fit statistics less than 1 tend to overfit to this model. Again, this result suggests important follow up by the investigators. While these raters do not show evidence of significant drift, the rating pattern here suggests a possible central tendency effect due to the raters' frequent application of "2" and "3" ratings across items. Such an effect may result from the raters' failure to distinguish between conceptually distinct and independent aspects of a teacher's practice as measured by a given item and again motivates a critical review of training procedures.

Insights into inter-rater reliability

Analysis of the model parameter estimates in Table 2 also provides insight into inter-rater reliability for this dataset. The rater bias facet shows that statistically significant differences among the raters ($p < .001$, $\chi 2 = 629.8$, df = 5) with a reliability of separation (reliability = .99). As above, as this index approaches 0, it indicates raters are interchangeable "scoring machines" that show no variance in their scoring practices (Eckes, 2006). Here, the separation approaches its maximum of 1, which indicates the raters are acting mostly independently with respect to their scoring practices. Moreover, the separation ratio is stratified into approximately 14 statistically distinct bins in terms of rating bias (strata = 14.39). Importantly, the model shows no significant misfit for any rater: despite the differences in bias and scoring practices, the raters have all produced ratings that are consistent with the expect scoring procedure. This interpretation is further supported by comparing the observed proportion of exact agreement among raters with the expected proportion of exact agreement. From the output table, the overall observed proportion of exact agreements (Obs% = 36%) is approximately close to the expected proportion of exact agreement (Exp% = 35.7%). With the marginally smaller Obs% than the Exp%, we can further strengthen the claim that the raters acted as independent experts in this study with distinct, yet valid codes.

These results offer the most compelling practical insights for investigators using MFRM to analyze teacher observation protocols: *when rater bias parameter estimates fall within acceptable ranges, individual rater scores can be adjusted for bias and included in overall estimates of teacher practices regardless of IRR metrics*. Here, the raters showed acceptable rating patterns such that the individual level proportional agreement statistics and fair average rating reported can be used to provide diagnostic information on IRR. The last two columns in Table 2 illustrate this information as the rater's observed scores (raw scores) and the rater's scores adjusted for rater bias (fair score). As is shown, Rater 1 and Rater 5 received similar observed scores in their rating, but their weighted fair scores differed drastically. By using the adjusted scores to support claims from the observation protocols, investigators can account for the distinct scoring practices of individual raters rather than exclude these differences in an effort to achieve homogenous scores among raters. This provides investigators with a powerful approach that leverages the differences among raters to achieve a more precise estimate of teacher practices that takes into account multiple rater analyses.

Discussion

Our results suggest that MFRM is a productive tool for identifying the quality of rater training programs as well as accounting for the variability in rater scoring. Using this approach, we analyzed the dataset produced from a project on teacher practice to assess the utility of MFRM for identifying rater outliers and for including rater variability in estimates of teacher practice. The resultant model demonstrated that the project's training methods

enabled most raters to consistently apply the rating scale over time with consistent bias; however, two raters exhibited inconsistent rating patterns across items. Our analyses showed these raters had low internal consistency despite continuous calibration training and suggests that codes from these researchers may require re-coding by another rater. Alternative methods, such as calculating Cohen's kappa using aggregated scores from the protocol—a common practice (Herrington, Yezierski, & Bancroft, 2016)—would have obscured these differences in item-level variability among raters. Without MFRM, these two raters' scores, which appear to demonstrate misunderstandings of the observed constructs assessed by the RTOP, may have compromised the validity of subsequent analyses of teacher practices.

Looking forward, MFRM could be used to identify raters with low internal consistency earlier in the rating process in order to administer additional training. Additionally, it can be used during analysis to adjust scores made by overly lenient or severe raters to more accurately measure teacher practice or to bring raters' scores within a benchmark range, such as the 0.8 kappa value suggested by Cohen (1960) and Krippendorff (1980). Most importantly, the kappa statistics are limited to measure how closely two raters agreed with each other, regardless of the reliability and validity of their coding practices. MFRM addresses this limitation by accounting for variability among raters that may result from individual rater expertise and experience to create a richer picture of changes in teacher practice. It would also be possible to use MFRM to identify items or raters as well as validity threats from halo effects or central tendency and account for them in the final model. In sum, MFRM provides an innovative way to support the scoring of teacher observation protocols by identifying and accounting for traditional issues with intra- and inter-rater reliability without compromising the individual perspectives of the raters.

References

Barbot, B., Tan, M., Randi, J., Santa-Donato, G., & Grigorenko, E. L. (2012). Essential skills for creative writing: Integrating multiple domain-specific perspectives. *Thinking Skills and Creativity*, *7*(3), 209-223.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, MJ: Lawrence Erlbaum.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education & Psychological Measurement*, *20*(1), 37-46.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Frankfurt: Peter Lang.

Fishman, B. J., Davis, E. A., & Chan, C. K. (2014). A learning sciences perspective on teacher learning research. In R. K. Sawyer (Eds.), *The Cambridge Handbook of the Learning Sciences* (pp. 707-725). Cambridge: Cambridge University Press.

Gomez, K., Kyza, E. A., & Mancevice, N. (2018). Participatory design and the learning sciences. In F. Fischer, C. E. Hemlo-Silver, S. R. Goldman, & P. Reiman (Eds.), *International Handbook of the Learning Sciences* (pp. 401-409). New York: Routledge.

Herrington, D. G., Yezierski, E. J., & Bancroft, S. F (2016). Tool trouble: Challenges with using self-report data to evaluate long-term chemistry teacher professional development. *Journal of Research in Science Teaching*, *53*(7), 1055-1081.

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: an example from the golden state examination. *Journal of Educational Measurement, 38*(2), 121-145.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: a meta-analysis. *Psychological Methods, 4*(4), 403-424.

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2019). Developing an explicit instruction special education teacher observation rubric. *The Journal of Special Education, 53*(1), 28-40.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment, 24*(2), 91-118.

Kaliski, P. K., Wind, S. A., Engelhard Jr, G., Morgan, D. L., Plake, B. S., & Reshetar, R. A. (2013). Using the many-faceted Rasch model to evaluate standard setting judgments: an illustration with the advanced placement environmental science exam. *Educational and Psychological Measurement*, *73*(3), 386-411.

Kramer, J., Kielhofner, G., Lee, S. W., Ashpole, E., & Castle, L. (2009). Utility of the Model of Human Occupation Screening Tool for detecting client change. *Occupational Therapy in Mental Health*, *25*(2), 181-191.

Krippendorff, K. (1980). *Content analysis*. Beverly Hills. California: Sage Publications, 7, l-84.

Linacre, J. M. (2019). *Facets Rasch model computer program*. Chicago: Winsteps.com.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*(1), 54–71.
Mannarini, S. (2009). A method for the definition of a self-awareness behavior dimension with clinical subjects: A latent trait analysis. *Behavior Research Methods*, *41*(4), 1029-1037.
McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555-576.
Myford, C., & Wolfe, E. W. (2009). Monitoring rater performance over time: a framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371-389.
Myford, C., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.
Myford, C., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.
Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed teaching observation protocol (RTOP) reference manual*. Tempe, Arizona: Arizona Collaborative for Excellence in the Preparation of Teachers.
Stieff, M., Nighelli, T., Yip, J., Ryan, S., & Berry, A. (2012). *The Connected Chemistry Curriculum (Vols. 1-9)*. University of Illinois: Chicago.
Toffoli, S. F. L., de Andrade, D. F., & Bornia, A. C. (2016). Evaluation of open items using the many-facet Rasch model. *Journal of Applied Statistics*, *43*(2), 299-316.
Wright, B. D., (2002). Number of person or item strata. *Rasch Measurement Transactions, 16(3)*, 888.
Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. Archives of physical medicine and rehabilitation. *Archives of Physical Medicine & Rehabilitation*, *70*(12), 857-860.

Acknowledgments

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170074 to the University of Illinois-Chicago. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. The authors express their appreciation to the science teachers and raters who participated on this project.