# SCORING WITH CLASSROOM OBSERVATIONAL RUBRICS: A LONGITUDINAL EXAMINATION OF RATERS' RESPONSES AND PERSPECTIVES

Temple A. Walkowiak
North Carolina State University
tawalkow@ncsu.edu

Jonee Wilson
North Carolina State University
jwilson9@ncsu.edu

Elizabeth Adams
Southern Methodist University
eladams@smu.edu

Anne Garrison Wilhelm
Southern Methodist University
awilhelm@smu.edu

*This study examines the utilization of cognitive interviews longitudinally over a one-year period to collectively trace raters' response processes as they interpreted and scored with observational rubrics designed to measure teaching practices that promote equity and access in elementary and middle school mathematics classrooms. We draw on four rounds of cognitive interviews (totaling 14 interviews) that involved four raters at purposeful time points spread over the year. Findings reported in this study focus on raters' responses about one rubric, positioning students as competent. The findings point to the complexities of utilizing observational rubrics and the need to track response processes longitudinally at multiple time points during data collection in order to attend to rater calibration and the reliability and validity of resulting rubric scores.*

Keywords: instructional activities and practices; research methods; measurement

In the field of education, researchers and evaluators regularly develop rubrics to assess or measure a particular construct with the intent for trained raters to apply the rubrics and assign reliable scores so that valid inferences can be drawn from the resulting data. One such example is classroom observational rubrics designed to measure mathematics teaching practices (e.g., Boston, 2012; Walkowiak et al., 2014). However, we know that using classroom observational rubrics is a complex and intense endeavor, due to the many nuances that exist in classroom interactions and instruction. Therefore, attending to how raters interpret the rubrics and apply their interpretations to assign scores is a critical type of validity evidence. That is, the response processes of raters can be utilized to evaluate if raters interpret the rubrics and apply scores as intended. In this study, we attend to raters' response processes over time using rubrics designed to measure teaching practices that promote equity and access in mathematics classrooms. The significance of this study is twofold: (1) it illustrates the complexity of one component of an interpretation/use argument (IUA) (Kane, 2016), but with longitudinal data; and (2) it draws the field's attention to the importance of iteratively examining raters' interpretations of rubric language and levels. It is critical that classroom observational rubrics generate reliable scores from which we can make valid inferences.

## Background

The EAR-MI (Equity and Access Rubrics for Mathematics Instruction) is a set of classroom observation rubrics designed to focus on specific practices that support more equitable participation and access in mathematics classrooms. The EAR-MI began as a theoretically-derived and empirically validated set of instructional practices (Wilson et al., 2019). For this study, we focus on raters' interpretation of one of the instructional practices and its accompanying rubric, positioning students as competent.

When teachers position students as competent, they explicitly and publicly value, identify, and acknowledge the brilliance of their students, framing their actions and statements as intellectually valuable (Bartell, 2011). The positioning rubric emphasizes the extent to which teachers specify what students do that is productive *and* the extent to which they provide rationales as to why what was done was considered productive. Figure 1 illustrates how positioning may be elevated in a classroom, corresponding to the levels on the rubric.

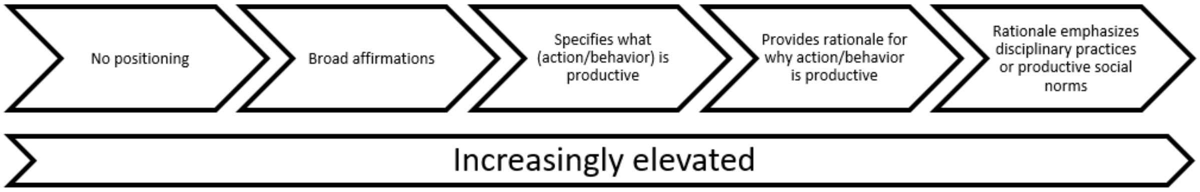
No positioning
Broad affirmations
Specifies what (action/behavior) is productive
Provides rationale for why action/behavior is productive
Rationale emphasizes disciplinary practices or productive social norms
Increasingly elevated

**Figure 1: Positioning students as competent, increasingly elevated.**

Our overarching goal in further developing the EAR-MI is to utilize Kane's (2016) "argument-based approach to validity" to systematically build an argument for validity. Kane describes the process of developing an interpretation/use argument (IUA) with claims to be evaluated with evidence. One such source of evidence is cognitive interviews to evaluate if the scoring is being applied accurately and consistently (Groves et al., 2009; Willis, 2005) and more specifically, to illuminate factors that could be impacting the scoring process. Cognitive interviews provide insights into four components of the raters' response processes: comprehension, retrieval, estimation, and scoring (Tourangeau, Rips, & Rasinski, 2000). Comprehension is the rater's process of understanding terms within a rubric and their combined meaning to interpret the rubric as intended. Retrieval refers to the rater recalling the necessary information or evidence in the video-recorded lesson. Estimation is the process of the rater judging the quality of the retrieved information for completeness and integrating information from their notes and memories to estimate a score. Scoring is the process of mapping the estimated response to the rubric's scale.  We utilized these four components in the context of the current study to longitudinally examine raters' interpretations of the positioning rubric.

## Methods

### Data Collection

Participants in this study were four raters on the scoring team for the EAR-MI. Each rater completed three or four cognitive interviews, conducted one-on-one with an interviewer. Before each cognitive interview, each rater watched a video-recorded mathematics lesson (the same lesson for all participants at the given time point). During the cognitive interview, the rater talked aloud about each rubric in light of the mathematics lesson, providing justifications and evidence for the scores they assigned.

Cognitive interviews occurred at four purposeful time points (TPs) across the course of one year. Four raters participated in the *first* interview, which took place in January 2021 (TP1), after a live, four-day rubric training, before raters entered a phase called training reliability. During the training reliability phase, raters scored lessons and subsequently met with an expert rater to discuss their scores. Two raters participated in the *second* interview, which occurred in April 2021 (TP2), mid-way through the training reliability phase. At this point, raters had scored 11 lessons and participated in an additional live training, what we refer to as "construct jams". During the construct jams, raters refined and adjusted their understanding of the rubric constructs. Four raters participated in the *third* interview, which occurred in June 2021 (TP3), after the training reliability phase and before raters scored lessons for a generalizability study. Following the training reliability phase, raters' scores indicated acceptable agreement with expert scores; 83% of raters' scores matched with expert scores exactly on the last five videos scored. At this point, raters had scored 10 additional lessons post- construct jams. Four raters participated in the *fourth* interview, which occurred in November 2021 (TP4), after the generalizability study and before raters started scoring lessons as a part of the larger sample of lessons to be included in continued examination of the validity of rubric scores. All interviews were transcribed. Having cognitive interview data at four time points allows for the tracing of interpretations of the rubric over time. It is important to note that in this study, we center raters' thoughts and perspectives in an effort to understand what is working and what needs improvement in scoring procedures as we aim to produce reliable rubric scores.

### Data Analysis

We focused on raters' collective interpretation of the focal rubric over the course of the four TPs. We chose the positioning rubric for two reasons. First, we are able to look across time at raters' interpretations of the rubric without changes in rubric language. This is possible because the generalizability study did not indicate that rubric changes were necessary. Second, the positioning rubric was functioning fairly well. For example, exact-match agreement rates between rater and expert scores exceeded 80%. We wanted to dive deeper with this rubric to understand the nuances that supported or hindered use from a rater perspective, particularly when everything looks like it is going well based on agreement statistics.

We reduced the 14 interview transcripts to only include when raters talked about scores for the positioning rubric. We conducted a line-by-line qualitative analysis of the raters' responses. After an initial read, we read through their responses using an open coding approach, tagging codes to the four components of the response process framework (Tourangeau et al., 2000). Codes were collapsed or fine-tuned based on a third reading of the data. We then identified themes for each component of the response process framework, focusing on the group of raters collectively, not an individual rater's development over time.

## Findings

Comprehension refers to raters' interpretations of the rubric's terms and their combined meaning. Across the first three time points, raters consistently grappled with the definition of a "rationale" and identifying when a rationale was present. As displayed in Figure 1, positioning is elevated when a teacher includes a rationale for why the student's action or idea is considered productive. In some instances, raters were "on the fence of whether or not [a teacher's comment] was a rationale" (TP3). Sometimes, the raters pondered the level of clarity and/or explicitness of the rationale (e.g., "I think it could be debatable because it's not 100% explicit, but this [teacher action] demonstrates the ability to provide rationales, but they are not necessarily clear" [TP3]). At TP4, the raters were much more decisive and did not grapple with the term as evidenced by "I'm going to stick with [my score] because I don't see a rationale."

Retrieval is the process of recalling the necessary information in relation to the rubric. Raters scored the lesson immediately after watching the lesson. They also utilized structured notetaking and applied "soft scoring" approximately every 20 minutes. "Soft scoring" means pausing and recording a score that represents what has happened so far in the lesson. The processes of structured notetaking and soft scoring were implemented between TP2 and TP3. Raters described how soft scoring "helps me calibrate and make sure that I know what I have strong evidence for" (TP3). Between TP1 and TP2, the positioning rubric changed from a rubric based on preponderance of evidence to highest evidence because the intent is to acknowledge a teacher's potential for implementing the equitable teaching practice of positioning students as competent. A direct implication of this change was evident in raters' ability to retrieve information for scoring efficiently, with less emphasis on finding every instance of positioning within a lesson: "I don't really have to search too far [in my notes] because I felt the earlier instance was stronger" (TP3).

Estimation occurs when the rater estimates a score based on notes and recollection of the video. With the exception of TP4, raters tended to ponder the distinctions between the levels on the rubrics. One rater described "the distinction between a two and a three is still kind of blurry….it would be hard for me to teach someone else what counts as a three. I'm not sure if I could articulate it clearly" (TP3). Here, the rater grappled with the levels on the rubric; this grappling corresponds to the uncertainty described earlier about the term, "rationale".

Scoring is the actual application of a score on the rubric. Across the four TPs, the raters became increasingly more confident, particularly at TP4 ("I'm always confident now"), when they spent less time grappling with the rubric and its terms, did not verbalize or demonstrate indecisiveness, and moved more quickly to assigning the score.

## Discussion

While our work is situated within the context of the EAR-MI, we present our discussion as three broader recommendations for researchers, both for those who are designing rubrics to measure a target construct and for those who utilize rubrics. Both groups should critically and carefully consider raters' interpretations and scoring applications. First, the cognitive interviews shed light on reasons for our raters' misinterpretations. If we did not have the cognitive interview data, our awareness of issues with the term, "rationale", would be less robust. Implementing cognitive interviews, whether in the context of the development of new rubrics or in the application of existing rubrics, is critical for understanding the nuances of raters' interpretations. Cognitive interviews also center the voices of raters and their perspectives about the important work they are doing. Second, we implemented procedures that resulted in better and more efficient retrieval of relevant information for scoring. The structured notetaking and soft scoring turned out to be fruitful as raters became more proficient with the rubric. When utilizing rubrics to score data, we recommend researchers attend to the process, not just the resulting scores. The systematic examination of these processes and their influence on raters seems to be an important component of collecting and evaluating evidence of validity. Finally, based on our read of the literature in mathematics education and beyond, cognitive interviews are typically implemented at one (or maybe two) TPs, and often only in the context of the initial development of a rubric. Our data suggests that iterative, longitudinal implementation of cognitive interviews is significant in improving raters' scoring procedures and thought processes. We were able to iteratively provide training to raters (as in the example of the "construct jams") throughout the year of these four rounds of cognitive interviews. In summary, the use of classroom observational rubrics to document and measure teaching practices is messy, but exciting work. Cognitive interviews at multiple time points serve as a mechanism for attending to the production of reliable scores such that valid inferences can be made when utilizing rubric data.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #1908481. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

## References

Bartell, T. (2011). Caring, race, culture, and power: A research synthesis toward supporting mathematics teachers in caring with awareness. *Journal of Urban Mathematics Education, 4*(1), 50-74.

Boston, M. (2012). Assessing instructional quality in mathematics. *Elementary School Journal, 113*(1), 76-104.

Groves, R. M., Fowler, F. J., Coupter, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley & Sons.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, United Kingdom: Cambridge University Press.

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & McCracken, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*(1), 109-128.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: SAGE.

Wilson, J., Nazemi, M., Jackson, K., & Wilhelm, A. G. (2019). Investigating teaching in conceptually-oriented mathematics classrooms characterized by African American student success. *Journal for Research in Mathematics Education, 50*(4), 362-400. https://doi.org/10.5951/jresematheduc.50.4.0362