TEACHER-RESPONSES: HIGHLIGHTING CHARACTERISTICS OF LOW RESPONSE PROCESS VALIDITY FOR ITEM(S) MEASURING TEACHERS' PEDAGOGICAL CONTENT KNOWLEDGE

Martha L. Epstein
UMass Dartmouth
mepstein1@umassd.edu

Hamza Malik
UMass Dartmouth
hmalik1@umassd.edu

Kun Wang
UMass Dartmouth
kwang1@umassd.edu

Chandra Hawley Orrill
UMass Dartmouth
corrill@umassd.edu

Response Process Validity (RPV) reflects the degree to which items are interpreted as intended by item developers. In this study, teacher responses to constructed response (CR) items to assess pedagogical content knowledge (PCK) of middle school mathematics teachers were evaluated to determine what types of teacher responses signaled weak RPV. We analyzed 38 CR pilot items on proportional reasoning across up to 13 middle school mathematics teachers per item. By coding teacher responses and using think-alouds, we found teachers' responses deemed indicative of low item RPV often had one of the following characteristics: vague answers, unanticipated assumptions, a focus on unintended topics, and paraphrasing. To develop a diverse pool of items with strong RPV, we suggest it is helpful to be aware of these symptoms, use them to consider how to improve items, and then revise and retest items accordingly.

Keywords: Item development; constructed response; pedagogical content knowledge; response process validity

Purpose of the Study

The mathematics education community has shown a growing interest in mathematics teacher knowledge assessment (e.g., Izsák et al., 2016 and Mosvold & Hoover, 2016). There is still much to do to understand the domain of mathematics teacher knowledge as well as to understand how to best assess this knowledge through a variety of types of scenarios and both CR (constructed response) and SR (selected response) types of items. When creating items for any assessment, one important task in which developers should engage is the determination of whether items exhibit response process validity (RPV). RPV is a measure of whether the person reading the items understands them in a way that is intended by the item developers. Our goal in this study was to determine the reasons PCK items may fail to demonstrate RPV. By being aware of symptoms of low-RPV, item developers can work to improve the assessment items, accordingly. Engaging in the RPV process can allow test developers to move beyond a binary decision to keep an item, or not, and may facilitate the development of assessment items with strong RPV that can also tap more complex, hard to communicate concepts that may take several rounds of revision to build strong RPV. We were guided by the following research question: What characteristics of middle school mathematics teachers' responses to CR PCK assessment items suggest low RPV?

Perspective and Review of Relevant Literature

Effective assessment development requires that relevant content domain (e.g., elements of proportional reasoning) first be successfully identified (Orrill & Cohen, 2016). For example, Hill and colleagues' (2008) efforts to measure mathematical knowledge for teaching, started by

defining the knowledge domain they hoped to measure. Defining the knowledge domain is challenging. Rowan et al. (2001) noted, "One difficulty we faced was developing items (and scenarios) that adequately tapped the full range of underlying "abilities" or "levels" of teachers' content and pedagogical knowledge…" (p. 16). Once the domain has been adequately conceptualized, the work of trying to measure that domain begins with the development of assessment items. Item development is complicated, in part, because the items must invoke the intended knowledge from test takers. This means that a participant needs to understand the question being asked in ways that align to the developer's intent. Therefore, it is important to determine the RPV of items. RPV has garnered recent attention in STEM education (e.g., Deng et al., 2021; Padilla & Benítez, 2014), and think-alouds is one method used to investigate RPV. Using methods such as think-alouds (Bostic, 2021), researchers sought to determine the match between test developer intent and test taker interpretation and to garner insights into what may be causing any mismatches (e.g., Mo et al., 2021). An area of opportunity is to continue to investigate sources and symptoms of low RPV and to work to hone our RPV research methods, so they lead to diverse items with strong RPV.

Methods

Context and Participants

Participants included a convenience sample of 13 middle school mathematics teachers from across the United States (nine female, four male). All the teachers were given pseudonyms. Assessment items used in this research study came from an assessment development effort seeking to measure mathematics teachers' content knowledge and PCK about proportional reasoning. The findings reported here are based on an analysis of only PCK items, all of which were constructed response. The items all involved asking teachers about realistic classroom scenarios (i.e., student work and classroom situations) and were based on Kersting and her colleagues' (Kersting, 2008; Kersting et al., 2012) successful work with similar items. Many of the items included short video clips that the teachers were asked to comment on.

Data Collection and Analysis

The 38 CR assessment items evaluated were spread across five surveys completed online. Depending on their schedule, teachers completed between one and five surveys and typed their responses to items. A follow-up Zoom think-aloud interview was recorded and transcribed. A coding template was developed to facilitate the analysis process that involved three researchers independently reviewing each item to determine, (1) if the item communicated as intended, (2) if the item did not communicate as intended, how had the teacher likely interpreted the item, and (3) if the item did not communicate as intended, characteristics of teachers' responses that led to this conclusion. Divergent views triggered a review of a teacher's data until the research team reached a consensus regarding how the assessment item "worked" for that teacher (e.g., did the teacher understand it as intended and did it invoke the kinds of reasoning intended).

Results

Teacher responses that created RPV issues tended to be: vague/overly general, predicated on assumptions, focused on unintended topics, and/or paraphrased information already provided. Below, we expand on our findings for each of these characteristics.

Vague, Overly General Answers

The intent of all CR PCK items was to elicit answers from teachers that provided rich insights into their PCK. Vague, overly general teacher answers to questions were, therefore the

antithesis of the types of answers item developers had hoped to solicit, thus, such answered were considered indicative of low item RPV. As an example of one vague response, participants were asked to watch a short video of a 7th-grade classroom discussing proportions. They were then asked, “Would you have led the class discussion in this video clip differently to support the student’s understanding of proportional relationships?” If “yes,” “explain how you would have led the discussion differently to support student’s understanding of proportional relationships.” Seven out of 10 teachers' responses to this item were coded as vague/overly general. For example, Christie offered, “…getting input for more than one student and perhaps leading them in a direction to the correct answer using keywords from students.” We considered this and other answers like it vague or general because Christie did not reference specific mathematical concepts, and while she mentioned she would look for "keywords," she did not provide detail about which key math words she was looking for or how she could help students make relevant connections among them to support an understanding of proportional relationships. We found vague/overly general responses were often associated with items that: were not specific enough, provided insufficient information, contained distracting elements, and/or did not adequately take the test-takers vantage point into account.

Answering Based on Assumptions

We considered another indicator of low item RPV when teacher responses suggested that teachers made assumptions to answer an item. Assumptions were deemed problematic because if test takers are making different assumptions, they are, in essence, answering different assessment questions. For example, for an item in which teachers were asked to watch a video and comment on a student’s understanding, Lydia, noted, “I think it was difficult to understand what he [the student] truly understood…” and Lydia went on to tell us that she “made an assumption about what he was sort of thinking …” In other instances, teacher responses suggested an assumption may have been inadvertently made which altered the question being asked. One such item presented an inverse proportion task (i.e., $y = k/x$), and the student’s use of cross multiplication to solve it—a strategy appropriate for solving proportional problems; not inverse proportions—was therefore problematic. Yet, seven out of 10 teachers’ answers were predicated on the assumption or assessment that the student’s work was correct, when item RPV depended on teachers recognizing the student’s work was not correct. In summary, when teacher responses explicitly or implicitly suggested an answer was based on an unanticipated assumption, we considered it a sign of poor RPV. We found teacher responses that were based on assumptions tend to be associated with items that provided insufficient information and/or did not ask specific enough questions.

Focusing on Unintended Topics

We deemed an item to have low RPV if teachers’ responses focused on an unintended topic, as this suggested they had interpreted the item quite differently than anticipated by item developers. An item in our research study that resulted in several responses that did not address the intended topic included a graph and table that featured a proportional relationship between mango weight and cost. The item stated, “how could you help students understand how the key characteristics of proportional relationships are demonstrated in both representations?”. This item’s goal was to solicit PCK regarding helping students build conceptual understanding of proportions across different types of representations. The item’s goal was not to tap PCK regarding optimal graph labeling. However, three of nine teachers’ answers focused solely on the suboptimal qualities of the graph. For example, Christie noted “you don't see any clear ordered pairs. If it was in units of one on the y axis, I think it would be easier for the students to visualize

the proportionality." Emma noted, "we can create ordered pairs: (1,6) (2,12), etc. Then I would place these points on the graph, so they can see that the points all fall on the line on the graph." In our study, we noticed teachers tended to have off-topic responses when an item contained a distracting element, was not specific enough, or did not adequately take the test-takers vantage point into account.

Paraphrasing the Information

Some teachers responded to items by simply paraphrasing information provided to them in the item scenario. No items were written asking participants to summarize the information provided. Instead, all items were designed to solicit PCK specific to a scenario. Hence, we considered paraphrasing a sign of RPV issues. In one example, teachers viewed a video clip in which a student (Evan) explained why he thought two ratios (6/14 and 15/35) were equivalent. Teachers were asked to "Comment on Evan's understanding based on his method, 'If you divide down, you'll get the same answer'." The intent was for teachers to use information about Evan and their own PCK to project what else Evan likely understood. Seven out of 10 teachers' responses did not go beyond paraphrasing Evan's response. For example, Christie noted that Evan "… understands that you can check for equivalent fractions by simplifying them." Similarly, Emma offered, "Evan understands that equivalent ratios will always simplify to the same numbers." In our study, paraphrasing was linked with overly general questions as well as items that did not adequately reflect the test-takers' perspective.

Discussion

Soliciting teacher feedback on assessment items via think-aloud or other relevant methodology is critical for developing items with high RPV. We posit that it is not only important to perform such RPV research to refine a given assessment, but also it is important to share "lessons learned" so that other research teams can benefit from better understanding potential pitfalls. Our findings are situated in one study, and we do not suggest the four response characteristics we found in our study are exhaustive of responses that signal RPV issues. We suggest identifying responses that signal low RPV is most beneficial if it is used to drive subsequent refinement of assessment items. We posit that using RPV as the basis for refinement may allow developers to create assessments that better measure challenging and harder to communicate ideas. We hope by sharing our analysis of "what teachers told us" in our assessment pilot items we will enhance existing knowledge for item development as well as trigger discussion regarding how one's RPV methodology may impact the diversity of assessment items that ultimately result.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1813760. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would also like to acknowledge Dr. Yasemin Copur-Gencturk for her support in providing teachers' dataset and other useful information for this research study.

References

Orrill, C. H., & Cohen, A. S. (2016). Why defining the construct matters: An examination of teacher knowledge using different lenses on one assessment. *The Mathematics Enthusiast*, *13*(1), 93-110.

Bonner, S., Chen, P., Jones, K., & Milonovich, B. (2021). Formative assessment of computational thinking: Cognitive and metacognitive processes. *Applied Measurement in Education, 34*(1), 27-45.

Bostic, J. D. (2021). Think alouds: Informing scholarship and broadening partnerships through assessment. *Applied Measurement in Education, 34*(1), 1-9.

de Bock, D., Van Dooren, W., Janssens, D., & Verschaffel, L. (2002). Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics, 50*(3), 311-334.

Deng, J. M., Streja, N., & Flynn, A. B. (2021). Response process validity evidence in chemistry education research. *Journal of Chemical Education, 98*(12), 3656-3666.

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W. D., & Rose, D. H. (2013). *A universal design for learning-based framework for designing accessible technology-enhanced assessments.* Pearson Assessment Research Report. http://images.pearsonclinical.com/images/tmrs/DolanUDL-TEAFramework_final3.pdf

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372-400.

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education, 20*(4), 427-441.

Izsák, A., Remillard, J. T., & Templin, J. (2016). Psychometric methods in mathematics education [Monograph]. *Journal for Research in Mathematics Education*.

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement, 68*(5), 845-861.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*(3), 568-589.

Mo, Y., Carney, M., Cavey, L., & Totorica, T. (2021). Using think-alouds for response process evidence of teacher attentiveness. *Applied Measurement in Education, 34*(1), 10-26. https://doi.org/10.1080/08957347.2020.1835910

Mosvold, R., & Hoover, M. (2016). *The Mathematics Enthusiast, 13*(1-2).

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144.

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., & Camburn, E. (2001). Measuring teachers' pedagogical content knowledge in surveys: An exploratory study. *Ann Arbor: Consortium for Policy Research in Education, University of Pennsylvania, 1*, 1-20.

Salkind, N. J. (2017). Tests & measurement for people who (think they) hate tests & measurement. Sage Publications.