# EARLY MATHEMATICS TEACHER PREPARATION EVALUATION RUBRICS FOR THE CONTEXT OF LIVE DISCUSSION FORUMS

Jeremy Zelkowski
The University of Alabama
jzelkowski@ua.edu

Tye Campbell
Crandall University
tye.campbell@crandallu.ca

*Teacher candidates bring many beliefs and interpretations of mathematics teaching and learning at the start of their teacher preparation coursework (e.g. methods courses, field experiences, assessment). Well-prepared beginning teachers in many instances requires programs and designed experience to breakdown unproductive beliefs and/or improve dispositions to align to best practices and equitable dispositions for the teaching and learning of mathematics. Our study focused on developing and validating two rubrics to evaluate teacher candidates' talk during live discussion forums in the first month of their initial teacher preparation program coursework with the intent to inform to varying degrees, where teachers candidate talk is situated (or not) in alignment to foundational readings and productive beliefs. Early validity evidence for rubric use is presented with suggestions for informative use and practice.*

Keywords: Affect, Emotion, Beliefs, and Attitudes; Preservice Teacher Education; Teacher Beliefs; Measurement

Many secondary mathematics teacher candidates (TCs) enter their preservice preparation programs with traditional (e.g. cultural norms, conventional) beliefs about education, believing, for example, that all or most student learn best in teacher-centered classroom instructional methods and/or performing rote practice with learn-by-example replication/practice (Cady et al., 2006; Conner et al., 2011; Cooney, 1999). Combating such beliefs is laborious since literature suggests TCs are likely to replicate practices most prevalent in their own educational experiences in K-12 mathematics classrooms (Borko at al., 1992; Cross, 2009). This idea is still generally supported by the more recent publication of the National Council of Teachers of Mathematics' (NCTM) *Catalyzing Change in High School Mathematics* (NCTM, 2017), where it is noted that only pockets of excellence exist within the high school for classrooms modeling the eight effective mathematics teaching practices (MTPs) (NCTM, 2014) developed from research over the last few decades. More troublesome in such conventional and culturally/historically normed mathematics teaching methods is the rooting of inequitable opportunities for non-dominant cultures creating a group of students becoming historically marginalized as an outcome (Aguirre et al., 2013). Our work is designed to fit within the challenges of preservice mathematics teacher education that presses to create critical dissonance within the status quo of teacher preparation programs with the designed intent of a resonant harmony of well-prepared, equitable teaching by newly minted mathematics teachers. We recognize this challenging work to rock the teacher preparation magnate model that has generally continued the cyclic pattern of student to TC to beginning teacher that does not significantly alter on a grander scale the teaching and learning of mathematics embedded in the eight MTPs.

Over the course of the last two decades, our preparation program has used and/or developed program measures as a means and intent to be predictive, yet deeply informative, with respect to desired outcomes for interns' (full-time student teachers) mathematics teaching practices and ultimately first-year teaching. We view the importance of internal program measures as needing

validity evidence for their effective use to inform and shape the development of all teacher candidates over time that helps shifts beliefs and practices when needed to desired TC outcomes. While we recognize the limitations of the data collection and rubrics presented in this paper, we believe this work presses preparation programs to push boundaries and challenge barriers that have not moved quickly enough to eradicate conventional and culturally/historically normed practices in mathematics teacher preparation to ultimately produce systematic change in K-12 schools in the teaching of mathematics to be equitable for all students. Our work firmly fits within most guiding questions for PME-NA 2022 program, as well as adding to the psychology of how the field views mathematics teacher preparation program components as a learning to teach experience that prepares and aligns to best practices as a systematic requirement to become a licensed teacher of mathematics. Our research questions are as follows:

1. What is the predictive nature of two rubrics scoring first semester teacher candidate discussion forums in relation to entering the teaching profession as validity evidence?
2. To what degree does the alignment of teacher candidates' first semester talk with assigned readings relate to completion of their preparation program?
3. To what degree do teacher candidates publicly stated first semester beliefs (productive or unproductive) during discussions relate to completion of their preparation program?
4. Do males and females (as identified on self-report) show any differences within the analyses given TC imbalance of 2:1 female to male within the preparation program?

## Theoretical Framework

The American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education *Standards for Educational and Psychological Measurement in Education* (AERA, APA, & NCME, 2014) introduces the need for validity evidence for measurement within educational contexts. The validation framework for the development, use, and interpretation of rubrics is the central tenet of our study.

Rubric development for affective domains in teacher education is an arduous task. Constructing a case for rubric validity is an ongoing process involving multiple facets of validity evidence, including the notion of the construct(s) in which are sought to be determined. Instruments in social sciences vary in precision and quality in the measurement of a particular construct, whereas the validation process should be centered on rubric interpretation and use (AERA et al., 2014; Bostic et al., 2019; Lavery et al., 2019). Kane (2013) denotes the use and interpretations of instruments [rubrics] require more validity evidence than is the case for less ambitious applications of instruments [rubrics]. Kane (2016) later indicated validation research is not easy but that it is generally sensible to grow solid evidence with manageable efforts.

### Situational Perspectives of Rubric Use in Teacher Preparation

Rubrics are used extensively in teacher education preparation programs for assessment and for accreditation purposes, but there has been a lack of validity evidence (published) with arguments for specific rubrics and their appropriate interpretation and use (Hill & Shih, 2009; Howell et al, 2019; Lavery et al, 2019). Well-constructed, systematically developed instruments with validity evidence in mind for their intended purposes have strong potential to provide teacher educators valuable and predictive data in the development trajectory of TCs. Traditionally, rubrics used programmatically are grounded in the documentation for accreditation purposes. That is to say for example, that for the Council for the Accreditation of Educator Preparation (CAEP) or jurisdictional level accreditation, rubrics are used to report aggregate level performance of program completers on an annual basis at different points in time within TCs' program

coursework and field experiences. Yet, most of the rubric data is not generate with a TC-level predictive nature in mind but rather for program accountability and documentation to become/remain accredited (CAEP, 2013, 2022). We challenge this status quo in teacher preparation in the use of rubrics beyond that of accountability, but significantly more so, with predictive validity in mind as a means to inform program faculty on their effectiveness of shaping TCs over time to align practices and beliefs to the mathematics education literature that ultimately removes barriers for marginalized students' opportunities to engage with and learning mathematics. Should such internal program rubrics exist, ultimately program faculty can modify program curricular experiences and provide TCs feedback identified earlier. We recognize that some TCs have only a single semester of preparation before the full-time (or extensive) student teaching internship. This fact is presented in our discussions and limitations later.

**Teacher Candidate Beliefs in Teacher Preparation**

Of particular importance to this study is NCTM's (2014) *Principles to Actions*, designed to describe "the conditions, structures, and policies that must exist for all students to learn" (p. vii). *Principles to Actions* (PtA) advances and redefines NCTM's (2000) six guiding principles for school mathematics: teaching and learning, access and equity, curriculum, tools and technology, assessment, and professionalism. Within each of these domains, NCTM (2014) outlined productive and unproductive beliefs that support or hinder mathematical learning for all students. NCTM's delineation between productive and unproductive beliefs in *PtA* serves as a standard by which TCs begin to think about their beliefs and how potential practices as a result of such beliefs may not be in the best interest of teaching and learning for all students.

Research suggests whole class discussions can positively influence TC beliefs within a whole class (Stohlmann, 2015; White et al., 2016) though less is known about the variation of influence and change on an individual TC level. The field would benefit from empirical research investigating the nature of individual contributions during whole class discussion and how those contributions shift beliefs toward or align to productive visions of teaching and learning mathematics within a teacher preparation programs. For example, if individual contributions during whole class discussion significantly influence TCs' teaching vision and beliefs, program faculty may utilize whole class discussions to learn about TCs' situated beliefs early enough to create situational interventions and even potential remediation activities. Stohlmann (2015) and White et al. (2016) both found engaging TCs in whole class discussions with purposefully chosen questions/tasks/prompts aided in TCs developing productive beliefs. Utilizing whole class discussion allows TCs to construct their own knowledge through discussing relevant material, but it further reinforces positive dispositions by enculturating TCs into a desirable learning environment. Our study is grounded in these ideas about using whole class discussions, which we refer to as TC live discussion forums, as the setting for the purposeful development of two rubrics to record and monitor TCs contributions over three live discussion forums in the first month of teacher preparation [education] coursework.

## Methodology

This paper reports on three distinct phases of the study in which rubric development was sought to create a tool to more reliably assess what teacher candidates put forward during live discussion forums. First, the development of two rubrics was needed to align to our research questions. Second, our team then looked to start the validity and revision process of testing the rubrics to move forward to the third phase. The final phase was to analyze the data in which the rubrics development sought to examine TCs alignment to the readings and productive beliefs (or not). Consider that in all mathematics teacher preparation programs, TCs who begin coursework

all do not finish the program as certified mathematics teachers. Therefore, our study methodology was to develop rubrics, test the rubrics on existing data to generate validity evidence, and then ultimately have validated rubrics situated to have a predictive ability to improve TC outcomes and program completion aligned to the desired characteristics in the field of mathematics teachers who exemplify productive visions and beliefs about teaching and learning mathematics (i.e. implementing the MTPs).

**Context of Participant Generated Discussion Forum Data**

Within the teacher preparation program at The University of Alabama, three live discussion forums have existed in the first month of the program for more than a decade. The notes about TC talk have been analyzed qualitatively in the assessment of progress in the program, but never has such analyses resulted in a mechanism to understand which TCs might program faculty provided deeper feedback on their development visions and beliefs about teaching and learning mathematics at such an early stage of the program. A seven-year period of data collection was used in which all three live discussion forums were implemented without deviation, holding the format, readings, and prompts to be consistent, as well as the mechanism for how data was noted and recorded in a spreadsheet. A small subset (~10%) via random selection of the existing data across all seven years was used for the development, test, revision, and initial validation work to finalize the rubrics for the full analysis. It is important to paint a vivid picture of how the data is collected during the live discussion forums.

The data used in this study was collected in a first semester TPP course (fall semester, junior year) which is embedded in a sequence of three consecutive mathematics teaching methods courses prior to the student teaching internship (see Zelkowski et al, 2018, 2021) for more details of program structure and sequence. The three live discussion forums took place at the start of the mathematics technology methods course during the first four weeks where each class was held twice a week for 1-hour 15-minutes (class meeting 2, 4, and 7). Lessons in the methods course between each of the three discussion forums included technology driven activities (see Zelkowski (2013), Bismarck et al., (2014) for examples) with TCs exposed as engaged learners to the eight Standards for Mathematical Practice (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010).

**Data Collection**

Over seven years, 110 PSMTs (77-female, 33-male) were enrolled in the methods course.

The readings were the first three chapters of *Teaching Secondary and Middle School Mathematics* (Brahier, 2013) titled: (a) Mathematics as a Process, (b) Principles of Mathematics Education, and (c) Learning Theories and Psychology in Mathematics Education. During each live discussion forum, the methods course professor projected a PowerPoint presentation with discussion prompts based on the focus questions of each chapter. The professor remained silent concentrating on accurate notetaking, allowing TCs to engage in an open, student-led forum related to each discussion prompt based upon the focus questions of each chapter. There were no restraints regarding how much individual TCs could talk during each prompt or discussion forum, though the professor encouraged equitable opportunities for all and respect among peers.

For this particular method of data collection (i.e. researcher as complete observer), Creswell (2003) discusses many advantages to this design such as it is useful in exploring topics that may be uncomfortable for participants to discuss, information can be discovered as it is revealed, and unusual aspects can be noticed with the researcher's firsthand experience with the participants. Creswell points to some limitations such as the researcher being intrusive or lacking good attention and observation skills. To mitigate such concerns, the methods professor had already

spent five years observing and recording data in such forums. Meetings with students allowed for the methods professor to check and verify the notes taken in years prior to the data used in this study. Given the five years of data collection and checking with students, we are confident that most comments captured reflect the TCs spoken words during the seven-year period of this study's data collection.

The professor took notes by constructing a grid with each TCs name as one column with additional columns for each prompt. TCs had name-tents on display for complete accuracy of comments being recorded. The quality of discourse and level of contribution during each of the three live discussion forums constituted about 3% of the final course grade (about 9% total) as an encouragement to participate early with peers at the start of the methods course sequence.

**Rubric Development Process**

The development of two rubrics and the validation work for their use to evaluate TCs' discussion contribution were rating two constructs of interest. The first rubric was designed to capture the alignment and interpretation of TCs' responses with the assigned chapter readings. The second rubric was designed to capture the alignment and interpretation of the responses to the unproductive and productive "beliefs about teaching and learning mathematics" within *Principles to Actions (PtA)* (NCTM, 2014, p.11) prior to TCs reading *PtA*.

We initially began with a four-level rubric (0-3) and two raters. We randomly scored about 3% of the forum data using the first rubric iteration. We discovered that an additional rubric-level would be needed and that a revision of the language was necessary as there were too many examples of implicit scoring decisions (e.g. higher or lower with less agreement). The rubrics were revised and retested with a new random sample of responses. After a third small revision of language in a few cells of the rubric, the third iteration was finalized.

In terms of each rubric, our revision process noted above went through three cycles in scoring a small sample until we were confident of removing as much subjectivity as possible between each level of the rubric. Ultimately, having enough levels in Rubric-1 was needed to consider the follow-up talk of a TC in terms of whether they were giving opinion supported by the interpretation of the reading or just giving opinion not supported by the reading. Whereas, Rubric-2, the sole purpose was to identify statements that do not align or do align to the productive beliefs tables in *PtA*. The rubric final iterations are shown in Figure 1.

**Rubric-1.** Alignment of Response with Reading

| Level 0 | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| Student did not engage in the discussion forum | Responses across prompts are mostly unrelated to the reading OR reveals misinterpretation of content of chapter | Responses across prompts are mostly related to the reading but is based on opinion rather than the content of the chapter | Responses across prompts mix personal opinion and correct interpretation of content of the chapter | Responses across prompts rely on content of the chapter but might also include some personal interpretation |

**Rubric-2.** Alignment of Response with *PtA* Beliefs about Teaching & Learning Mathematics

| Level 0 | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| Student did not engage in the | Response across prompts reveals statements which align with | Response across prompts partially reveals statements which align with at | Response across prompts reveals more statements which | Response across the prompts reveals little evidence of statements that align |

| | | | | |
|---|---|---|---|---|
| discussion forum | unproductive beliefs | least one productive belief | align with productive beliefs than not | with unproductive beliefs |

**Figure 1: Finalized Rubrics for Scoring TCs Alignments within the Discussion Forums**

## Results

### Rater Reliability of Rubrics

One rater scored all three discussion forums for all N=110 TCs with both rubrics which generated 660 spreadsheet cells. We randomly selected 10% of the cells for each rubric for a second rater to score independently. The rater agreement was 78.8%. We discussed our scores that were not in alignment in reference to the rubrics and the TCs discussion forum contributions. Rater one revisited all 660 cells and rescored. Independently, rater two randomly selected another 10% of cells to score for each rubric. There were 6 cells of overlap with the initial 10% selected for a total of 126 cells (19.1% of all). Rater two rescored the initial 66 cells and scored the additional 60 cells. The rater agreement was 88.1%. The results of each rubric scoring are presented in Table 1. We proceeded then to compute the associated Cohen's Kappa, Cronbach Alpha, and Intra-class Correlation Coefficient (ICC) presented in the note of Table 1.

**Table 1: Rater Agreement for Rubric-1 and Rubric-2**

| Rubric-1 | Rater 2 | | | | | Rubric-2 | Rater 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rater 1 | 0 | 1 | 2 | 3 | 4 | Rater 1 | 0 | 1 | 2 | 3 | 4 |
| 0 | 6 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 9 | 3 | 0 | 0 |
| 2 | 0 | 0 | 29 | 4 | 0 | 2 | 0 | 0 | 19 | 2 | 0 |
| 3 | 0 | 0 | 1 | 19 | 0 | 3 | 0 | 0 | 1 | 20 | 1 |
| 4 | 0 | 0 | 0 | 2 | 1 | 4 | 0 | 0 | 0 | 1 | 1 |

*Note.* Rubric-1 Statistics: Kappa=0.819, ICC=0.934, Cronbach Alpha=0.966. Rubric-2 Statistics: Kappa=0.822, ICC=0.939, Cronbach Alpha=0.969. All statistics sig. @ $p<0.001$ level.

### Relationship of Rubric Scores to Program Completion

Rubric-1 and Rubric-2 produce two scores for each of the three live discussion forums and total scores (sums of each forum). Of the N=110 TCs, 64 were program completers (~58%), 39 left the program (~35%) for academic and/or personal decisions not to enter teaching, and seven (~6%) eventually entered teaching through an alternative non-certified route. With this knowledge, our interest is whether Rubric-1 and/or Rubric-2 have any predictiveness in classifying TCs as a likely program completer or not from semester-1 discussions. More importantly, rubric scores should provide methods faculty early insights to improve outcomes related to core program readings, beliefs, and ultimately practices.

Two quantitative analysis methods were employed to address the research questions. Zelkowski (2011) stated, "the basic idea underlying this statistical method was to determine whether these groups differed significantly with respect to the mean of" (p.34) rubric levels individually and collectively. Zelkowski further described the use of Discriminant Analysis as a method that indicates the nature of the predictor variables (i.e. both rubrics) in contributing to group separation (i.e. outcome group) through a linear modeling process. We used Chi-square for each rubric across each of the three live discussion forums (six total analyses) considering our TCs three possible outcomes. Because research suggests males and females contribute differently

to whole class discussions (e.g. Guzzetti & Williams, 1996), we further examined if there were any differences across TCs gender.

**Statistically Significant Findings**

Rubric-1 was found to be least predictive of program outcome for all three forums. However, the Chi-Square analyses found statistically significant results for Rubric-1 with the third discussion forum chapter (Learning Theories and Psychology in Mathematics Education) for females but not males. Given the ratio of 77:33 TCs' gender imbalance, this is not surprising. This is further examined in the discussion.

The strongest results were found with the Discriminant Analyses. We found correctly classify males was significant in predicting a final outcome, but not females alone and not all N=110 TCs together in the analyses. Based on each Rubric-2, as well as the sum of both rubrics, males were correctly classified in the analyses into their outcome between 60% and 76%. Rubric-2 scores for chapter 1 forum predicted the outcome 60.6% correctly, chapter 2 forum predicted the outcome 69.7% correctly, and the sum of all three forums predicted the outcome 75.8% correctly. When considering both rubric scores on all three forums (six total scores summed), male outcomes in the program were predicted 69.7% correctly.

## Discussion

The purposes of the discussion forums have always been driven by introducing TCs to three foundational chapters setting the stage for mathematics teaching and learning. More importantly, the design engages TCs with peers to hear, listen, and discuss with each other's interpretations of the readings and injecting their own beliefs. Because these three forums immediately provide about 9% of the course grade, there is an early responsibility of TCs to read, be ready to discuss, and openly discuss critical foundations of mathematics education with their peers. We report here that less than 5% of TCs did not engage at least in at least one of three discussion forums.

**Program Use of Rubrics**

Our intent in the development of these rubrics and working through validation processes for their use and interpretation with desired outcomes, was and still is, to understand TCs potential cognitive conflicts with aspects of mathematics teaching and learning best practices. We further wanted to understand if early semester-1 discussions could be assessed to improve attrition and begin building improved belief structures. Our use of these rubrics in the program serves as informative early indicators of cognitive conflict with readings and/or productive beliefs about teaching and learning mathematics. In our previous program analyses (Zelkowski et al., 2018, 2021), we focused on the impact of key program assessments, coursework, and structural sequence on content knowledge and pedagogical content knowledge for program completers. These two rubrics across three live TC discussion forums provides an early indicator in which personal emails, one-on-one conversations, and high-quality feedback can be provided to TCs to increase their likelihood of program completion if teaching is truly their career desire. We do not see these rubrics as valuable in saving or rescuing those who decide to change majors out of teaching, though we do see their use as indicative and valid to improve outcomes and reduce attrition.

**Implications for Mathematics Teacher Education**

Rubrics are widely used in teacher education as we previously discussed, including mathematics teacher education. Rubrics serve excellent purposes regarding accreditation of programs, admission to teacher education programs, assignments and grades in coursework, and to provide feedback to TCs. However, the literature rarely points to the issues with the validity of using

self-created rubrics for assessment without rigorous validity work having ensued (Hill & Shih, 2009; Howell et al., 2019; Lavery et al., 2019). That is to say, such rubrics generally lack one or more of the categories of validity evidence, or have none at all, for establishing arguments for rubric use in making important interpretations and decisions in mathematics teacher preparation. More precisely, what constructs do rubrics measure and to what degree are the use and interpretation valid? These are difficult questions to face without some process in the validation of the rubrics that methods course faculty may use. We provide our results and findings as a way to encourage rubric use but by generating validity evidence for use and their interpretation. Rubrics as any sort of measure in mathematics teacher education programs, ultimately should be providing faculty and TCs some predictive value about their development towards first-year teaching. Further, we demonstrate a methodology that aligns to the AERA, APA, and NCME (2014) for using rubrics for purposes of decision making, assessment, and programmatic outcomes with aims to reduce attrition and provide objective scoring for TCs.

## Conclusion

As researchers, we are interested in examining the 'next step' of remediation when TCs do not exhibit movement towards productive beliefs regarding teaching and learning and/or share personal experiences as 'proof' over well-structured readings (i.e. Brahier). That is, how can methods faculty support learners whose beliefs regarding teaching and learning are not influenced by readings and discussions? We made several revisions to our program design over the last decade to account for these considerations (Zelkowski et al., 2018; 2021; Zelkowski & Gleason, 2018), but there is much to learn regarding effectiveness of program components, course designs, and the validation of key assessments.

We hope this work provides stimulus for discussion, research, and practice towards aiding TCs in developing productive beliefs about teaching and learning mathematics within their preparation program. Developing and validating rubric use and interpretation is a critical piece to not only improving mathematics teacher education, but also measuring the effectiveness of interventions aimed at stronger TC development. This work may serve as an example of the utility of rubrics in mathematics teacher education and provide methods faculty with a process for the validation of rubric use in critical junctions of teacher preparation programs.

## Acknowledgments

This work was supported in part by the National Science Foundation (NSF) Grants #1849948, 1340069. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the NSF.

## References

Bismarck, S., Zelkowski, J., & Gleason, J. (2014). Predicting future gas prices using the standards for mathematical practice. *Mathematics Teacher, 107*(9), 694-700.

Aguirre, J. M., Mayfield-Ingram, K., & Martin, D. B. (2013). *The Impact of Identity in K-8 Mathematics: Rethinking Equity-based Practices*. Reston, VA: National Council of Teachers of Mathematics.

Brahier, D. J. (2013). *Teaching secondary and middle school mathematics* (4th edition). Pearson.

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P. C. (1992). Learning to teach hard mathematics: Do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*(3), 194–222.

Bostic, J., Krupa, E., & Shih, J. (2019). Introductions: Aims and scope for assessments in mathematics education contexts: Theoretical frameworks and new directions. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 1-11). Routledge.

Cady, J., Meier, S. L., & Lubinski, C. A. (2006). The mathematical tale of two teachers: A longitudinal study relating

mathematics instructional practices to level of intellectual development. *Mathematics Education Research Journal, 18*(1), 3–26.

Conner, A., Edenfield, K.W., Gleason, B.W. & Ersoz, F.A. (2011). Impact of a content and methods course sequence on prospective secondary mathematics teachers' beliefs. *Journal of Mathematics Teacher Education, 14*(6), 483–504.

Cooney, T. J. (1999). Conceptualizing teachers' ways of knowing. *Educational Studies in Mathematics, 38*, 163–187.

Council for the Accreditation of Educator Preparation (CAEP). (2013). CAEP Initial Level Standards. Author.

Council for the Accreditation of Educator Preparation (CAEP). (2022). CAEP Initial Level Standards. Author.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approach* (2nd edition). Sage.

Cross, D. (2009). Alignment, cohesion, and change: Examining mathematics teachers' belief structures and their influence on instructional practices. *Journal of Mathematics Teacher Education, 12*, 325–346.

Guzzetti, B. J., & Williams, W. O. (1996). Gender, text, and discussion: Examining intellectual safety in the science classroom. *Journal of Research in Science Teaching, 33*(1), 5-20.

Hill, H., & Shih, J. (2009). Examining the quality of statistical mathematics education research. *Journal of Research in Mathematics Education, 40*(3), 241-250.

Howell, H., Stone, E., & Kane, M. (2019). Future directions in the measurement of mathematics teachers' competencies. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 230-252). Routledge.

Kane, M. (2013). Validation the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64 80). Routledge & Taylor & Francis Group.

Krupa, E., Bostic, J. & Shih, J. (2019). Validation in mathematics education: An introduction to quantitative measures of mathematical knowledge: Researching instruments and perspectives. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 1-13). Routledge.

Lavery, M., Jong, C., Krupa, E. & Bostic, J. (2019). Developing an instrument with validity in mind. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 12-39). Routledge.

National Council of Teachers of Mathematics. (2014). P*rinciples to actions: Ensuring mathematical success for all.* Author.

National Council of Teachers of Mathematics. (2017). *Catalyzing change in high school mathematics: Initiating critical conversations.* Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards Mathematics.* Author.

Stohlmann, M., Moore, T., Cramer, K., & Maiorca, C. (2015). Changing pre-service elementary teachers' beliefs about mathematical knowledge. *Mathematics Teacher Education and Development*, *16*(2), 4-24.

The American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Measurement in Education.* American Educational Research Association.

White, D. Y., DuCloux, K. K., Carreras-Jusino, Á. M., González, D. A., & Keels, K. (2016). Preparing preservice teachers for diverse mathematics classrooms through a cultural awareness unit. *Mathematics Teacher Educator*, *4*(2), 164-187.

Zelkowski, J. (2011). Defining the intensity of high school mathematics: Distinguishing the difference between college-ready and college-eligible students. *American Secondary Education, 39*(2), 27-49.

Zelkowski, J. (2013). Making sense of extraneous solutions. *Mathematics Teacher, 106*(6), 452-458.

Zelkowski, J., Campbell, T.G., & Gleason, J. (2018). Programmatic effects of capstone math content and math methods courses on teacher licensure exams. In Smith, W.M., Lawler, B.R., Strayer, J.F. & Augustyn, L. (Eds.). *Proceedings of the 7th Annual Mathematics Teacher Education – Partnership Conference* (pp. 91-96). Association of Public Land-grant Universities.

Zelkowski, J. & Gleason, J. (2018). Programmatic effects on high stakes measures in secondary math teacher preparation. In Venenciano, L. and Redmond-Sanogo, A. (Eds.). *Proceedings of the 45th Annual Meeting of the Research Council on Mathematics Learning.* Baton Rouge, LA.

Zelkowski, J., Campbell, T.G., Moldavan, A.M., & Gleason, J. (2021). Maximizing teacher candidate performances

based on internal program measures: Program design considerations. In Smith, W.M. & Augustyn, L. (Eds.). *Proceedings of the 10th Annual Mathematics Teacher Education – Partnership Conference* (pp. 39-42). Association of Public Land-grant Universities.