OPPORTUNITIES FOR MATHEMATICS ENGAGEMENT IN SECONDARY TEACHERS' PRACTICE: VALIDATING AN OBSERVATION TOOL

Amanda Jansen
University of Delaware
jansen@udel.edu

Ethan P. Smith
University of Delaware
etsmith@udel.edu

James A. Middleton
Arizona State University
jimbo@asu.edu

Catherine E. Cullicott
Arizona State University
ccullico@asu.edu

The purpose of this report is to present our process and results for establishing validity and reliability of an observation tool used to investigate teaching practices that high school mathematics teachers use to engage students. We developed our tool using established practices, such as reviewing literature to develop a framework for instruction and piloting the tool to design descriptive levels for rubrics. After validating externally by consulting experts, additional rubrics regarding teaching mathematics for equity were added to the tool. We conducted a reliability study of 149 episodes of classroom instruction (equivalent to 447 10-minute segments of instruction in all), two raters per episode, to investigate the nature of coding disagreements. Most disagreements occurred due to raters noticing different evidence rather than different interpretations of rubrics, which suggested the value of two raters and resolution meetings.

Keywords: Research Methods; Instructional Activities and Practices; High School Education; Affect, Emotion, Beliefs, and Attitudes

A range of observation tools exist to support the study of mathematics instruction that supports students' learning of mathematics (e.g., Bostic et al., 2019; Boston, 2012; Hill et al., 2012; Sawada et al., 2002; Walkowiak et al., 2014). These tools enable researchers to compare teaching practices, such as features of classroom discourse, that align with frameworks for quality instruction. Although these observation tools are well-established and validated, they focus primarily on behaviors that can explain students' learning, such as mathematical task enactment; they do not investigate how mathematics teaching influences students' engagement.

It is important to identify teaching practices that can motivate and engage students in mathematics classrooms, particularly in secondary grades. It has been well documented that students' mathematics engagement decreases over time as they move through levels of education into high school (e.g., Collie et al., 2019). Students' self-efficacy, enjoyment, and sense of the utility of mathematics tends to decrease as they move from elementary school into junior high (Wigfield et al., 1991); this trend continues through high school (Chouinard & Roy, 2008). However, students' motivation and engagement is socially situated and influenced by teachers' instructional practices in the moment (Anderson et al., 2004; Shernoff et al., 2017), so it is important to investigate teaching that supports engagement. The purpose of this paper is to describe the validity and reliability of the observation tool that we developed for the SMiLES project [Secondary Mathematics in-the-moment Longitudinal Engagement Study] to investigate how secondary mathematics teaching may impact students' engagement.

Potentially Engaging Mathematics Instructional Practices

For students to learn mathematics, they must be engaged. We conceptualize engagement in mathematics classrooms as a person's cognitive, affective, behavioral, or social investment in a pedagogically relevant object, such a mathematics task or lesson, as situated in the relationship between the self, the object of engagement, and others in the environment (Middleton, Jansen, & Goldin, 2017). In a study of almost 4,000 middle school and high school students in Western Pennsylvania, higher levels of cognitive, behavioral, emotional, and social engagement predicted students' course grades in mathematics (Wang et al., 2016). According to Greene (2015), it is well-established in prior research that motivation constructs such as students' self-efficacy support students' engagement in ways that lead to learning.

Instruction is likely to support students' engagement when teachers provide students with both *social support* for working together on content and *academic support* for accessing rigorous mathematical content (Shernoff et. al., 2016). Such support can take a variety of forms. *Academic support* may include opportunities for sense-making and reasoning (Stein et al., 1996); opportunities to make conceptual connections (Hiebert & Lefevre, 1986); pressing students to explain their thinking (Engle & Conant, 2002; Kazemi & Stipek, 2001); providing students with specific and detailed feedback (Stipek et al., 1998), opportunities to solve mathematics tasks in context (Koedinger & Nathan, 2004); or some combination of these. *Social support* may include motivational discourse with a focus on learning, positive affect, and encouragement of collaboration with peers (Turner et al., 2002); positioning students as competent (Cohen & Lotan, 1995; Gresalfi et al., 2009); accountability practices in the classroom (Horn, 2017); providing opportunities for student-to-student discourse in whole class discussions (Nathan & Knuth, 2003) or small groups (Fuentes, 2018) in ways that maintain mathematical quality; attention to students' lives outside of school (Yamauchi et al., 2005); or some combination of these teaching practices. Whether these supports can foster students' mathematical engagement remains at the level of conjecture, and an observation tool could explore this conjecture.

Development Process and Use of our Observation Tool

The SMiLES project's observation tool measures the extent to which potentially engaging teaching practices are present in a lesson. The tool does not establish whether instruction was engaging for students. Student engagement in observed lessons was assessed by an in-the-moment student survey using Experience Sampling Methodology (Jansen et al., 2019; Schiefele & Csikszentmihalyi, 1995).

The final version of the tool includes fifteen rubrics to assess eight dimensions of academic support and seven dimensions of social support. Rubrics designed for academic support measured students' opportunities for sense making and reasoning, connections between representations or strategies, pressing students to explain, contexts of tasks, mathematical correctness, mathematics language use, feedback, and students' opportunities for agency and autonomy. Social support rubrics assessed whole class discourse, small group discourse, status raising and positioning students as competent, motivational discourse, enthusiasm about mathematics, attention to students' lives, and accountability and high expectations. We defined each dimension with descriptive levels. Each dimension was scored on a four-point rubric with points (0-3) assigned to index each level: absence or the opposite of ideal enactment (0), weak level of enactment (1), moderate level of enactment (2), and strong level of enactment (3). Each rubric included a definition of the teaching practice, and we defined the observable indicators for each level of enactment. We share an example observation rubric below in Figure 1.

Social Support 6: Attention to Students' Lives

This rubric captures the degree to which the teacher attempts to connect with students' lives while teaching.

Strongly Present (3)	The teacher speaks about more than one example of cultural events or outside of school events during instruction OR talks with multiple students about aspects of their lives outside of school or mathematics class in ways that are *incorporated into instruction.*
Moderately Present (2)	The teacher speaks about one example of cultural events or outside of school events during instruction OR talks with one student about aspects of their lives outside of school or mathematics class in ways that are *incorporated into instruction.*
Minimally Present (1)	The teacher mentions cultural events, outside of school events, or other information personal to any students during class, but does not incorporate it into instruction.
Not Present (0)	The teacher does not mention/discuss anything personal to students during instruction.

Figure 1: Rubric for teachers' efforts to attend to students' lives

The SMiLES project's observation tool is designed to investigate potentially engaging teaching practices during an activity within a lesson. Before each classroom observation, we asked teachers to complete an online form in which they would nominate a potentially engaging activity that would take place within the lesson. Members of our research team video-recorded the entire class period, with a particular focus on these activities, which ranged from roughly 10 to 45 minutes with a median length of 30 minutes. These teacher-selected potentially engaging instructional activities lasted between 9 minutes and 40 seconds and 45 minutes and 40 seconds, with a median length of 29 minutes and 42 seconds.

We applied the observation tool rubrics to the video recorded activities in 10-minute segments. If the last segment was under three minutes, it was not rated. Each 10-minute segment in an episode was individually rated by two coders on the research team, who then met to resolve disagreements in scoring. Coders resolved disagreements in their segment ratings for each rubric by describing the observed behavior they used as evidence when scoring and how they interpreted that behavior within the framework of a rubric. Resolved scores for each segment were assigned. Episode scores were determined by averaging the resolved segment scores for each rubric.

To calibrate rating criteria and to address potential coding drift, all observation team members met at least once per academic semester to train for rating consistency with the observation tool and to resolve any outstanding questions that had arisen during the resolution procedures. Training involved all raters coding the same episode independently and meeting to resolve disagreements as a team. Orientation to the coding concepts also included reading and discussing relevant literature as a team (e.g., Middleton, Jansen, & Goldin, 2017).

Research questions. This report consists of two studies that respectively illustrate the validity and reliability processes used for the SMiLES project's observation tool. These studies answer two research questions:

1. Validity study: To what degree did the rubrics in the observation tool align with appropriate phenomena (instructional practices that promote mathematical engagement)?

2. Reliability study: To what extent did raters in our research team reach agreement when rating observation episodes? When there was initial disagreement, what explained lack of agreement?

Method: Participants and Context

The SMiLES project team collected classroom observation data from 29 secondary mathematics teachers' lessons in two U.S. states. Sixteen of these teachers taught in a mid-Atlantic state and 13 taught in a southwestern state. Twenty-one teachers were female and eight were male. The teachers also represented a variety of racial and ethnic backgrounds, with 22 teachers identifying as white, two identifying as Black, two identifying as Latinx, and one each identifying as Asian, Black/Hispanic, and White/Asian. The teachers worked with a diverse student population. In the Mid-Atlantic, the schools' demographics ranged from 12-34% low income, 25-60% white, 27-47% Black, and 6-21% Latinx. In the Southwest, the schools' demographics ranged from 76-94% low income, 1-6% white, 1-16% Black, and 77-96% Latinx. We targeted courses at the equivalent of on-grade level mathematics for ninth and tenth grade students, which included topics-based courses in the southwestern U.S. (Algebra I, Geometry) and integrated courses in the mid-Atlantic (Integrated Math [IM] 1, IM 2, IM 3). Each class period was observed two or three times during a course. A course was either one semester (if on block scheduling, such as schools in the mid-Atlantic) or a full academic year (southwest schools). The reliability study was conducted on a subset of these data.

Validity Study

Procedures

Characterizing teaching is a qualitative practice, and we conceptualize validity as multifaceted in qualitative work. Hayashi et al., (2019) present a variety of validity frameworks for qualitative work, including the following: *Descriptive validity* concerns the ability of the report of an event to faithfully record its important features. The interdependence of observations and the descriptions of those observations must be developed from theory. *Interpretive validity* concerns the ability of the tool to help the researcher construct the meaning of the events and the behaviors of the people engaged in those events. *Theoretical validity* refers to the consistency of the analytic coding and the theoretical argument that is constructed. It is thus concerned with the truth of the concepts and classifications developed in the analysis, and the ways in which the concepts and classifications interrelate in the abstraction of the event to the (nascent or developing) theory. *Validity generalization* refers to the ability of the method to be used in other situations, times, and places. For an instrument such ours, its descriptive and interpretive frameworks and its theoretical validity should be applicable in a new context.

Our first step toward internal conceptual validity was to operationalize the construct of engaging secondary mathematics instruction grounded in a theoretical frame from research literature (theoretical validity). This framework was developed by two researchers with expertise both in mathematics teaching and learning from mathematics education and motivation and engagement from educational psychology. For rubric development, we then translated the theoretical framework into descriptive rubric levels for each teaching practice. To internally examine construct validity in the rubrics, the entire research team (composed of graduate students and faculty with expertise in mathematics education or psychology) met multiple times to discuss whether and how these levels reflected the desired teaching practices and whether the descriptions were observable and amended accordingly.

We then piloted the tool by rating publicly available video from the TIMSS video study [http://www.timssvideo.com/] (descriptive validity). This pilot study involved all members of the research team rating the same two videos using the rubrics. The team met as a whole group to compare and contrast their ratings. Disagreements were discussed and the levels of enactment for each rubric were then specified further. A rubric to describe the nature of teacher feedback was added to reflect this teaching practice as a result of piloting.

To externally examine construct validity of the observation tool rubrics and individual rubric levels, we shared the observation tool with an expert panel, the SMiLES project's advisory board, which consisted of experts in educational psychology and mathematics education (interpretive and theoretical validity). All of the researchers in the advisory board had studied mathematics or science engagement in the context of learning environments, and they had all developed methods for studying teaching practices that support students' engagement. The results of the validity study reflected the team's learning from the expert panel.

Results

We initially generated twelve dimensions or rubrics based on our review of the research literature, and the expert review panel for our validity process led to three new rubrics. At the team's first advisory board meeting, three months into the three-year project, we shared the first draft of the observation codebook. The draft reflected a review of the literature, internal construct validity meetings, and revisions to the codebook after piloting it. Advisory board members then suggested additional rubrics that supported equity in mathematics teaching and learning.

As a result of external feedback, we revised the observation tool to reflect a broader conceptualization of equity in potentially engaging mathematics teaching (interpretive validity). In our initial rubrics, we approached equity primarily as access by writing rubrics that measured opportunities for students to experience sense making and reasoning, connections, tasks in context, and other aspects of high-quality mathematics instruction. We acknowledged that access is only one dimension of equity (Gutiérrez, 2002). Some of these rubrics were more closely aligned with supporting students' identities.

We added three rubrics related to promoting equity in mathematics teaching after feedback from our advisory board, resulting in 15 rubrics in all. We added a rubric about *attention to students' lives* in mathematics teaching (see Figure 1) (Yamauchi et al., 2005). Attention to students' lives could align with identities as students could begin to see themselves reflected in mathematics. We also added an *accountability* rubric to examine whether and how teachers held students to high expectations, acknowledging that high expectations are necessary but not sufficient to achieve equity (Lubienski, 2002). One final rubric was added after the pilot year of data analysis: *student enactments of agency and autonomy.* Our initial rubrics did not appear to capture the opportunities that students had to exhibit control over their own learning (Kosko, 2016). Opportunities to enact agency are chances for students to develop productive mathematics identities (Gresalfi et al., 2009).

Reliability Study

Data Sources

To investigate whether and how our rubrics could be applied consistently across raters, we conducted a reliability study of the analysis of 149 video-recorded episodes of teacher-selected potentially engaging activities. Each of these observation episodes was rated by two analysts. We dispersed resolution assignments to ensure that duplicate pairs of raters were minimized. Across the 11 raters who were on the team at any point from 2018-2020, 43 unique pairs of raters were assigned to resolve scores. The most resolutions shared by any single pairing was 19 episodes.

The resolution process began with both raters independently scoring each 10-minute segment of the observation video across each of the 15 rubrics in the observation tool. Every score was justified by documentation of evidence from the recorded observation, including timestamps. Raters then met to discuss any discrepancies in their initial ratings and to resolve the scores for each segment. This resulted in resolved scores for each rubric by segment, as well as episodic scores which were the average of resolved segment scores for each rubric. Every observation which was coded involved this resolution process; no episode was analyzed by a single rater.

During the final round of observation resolutions, raters also identified the nature of any initial disagreement in their individual scores to better understand the reliability of the observation tool. They identified whether any initial disagreement was the result of one rater noticing additional evidence in the video (resulting in a higher or lower score) or whether the initial disagreement was the result of conceptual differences between the raters with regards to the rubrics themselves (i.e., the same evidence was given different ratings).

Analysis Procedures

Following paired ratings, the Intraclass Correlation Coefficient (ICC) for each rubric was computed to examine the reliability of initial ratings *prior* to the resolution process. An excellent ICC, or a value greater than 0.9 (Portney & Watkins, 2000), could indicate that the resolution process was unnecessary (i.e., raters almost always agreed on the rubrics in their initial ratings). In contrast, a poor ICC of less than 0.5 (Portney & Watkins, 2000) suggests value in resolving.

The ICC used to determine reliability was a 2-way random effects model. Initial segment ratings were converted to a “low” and “high” score depending on how the two initial raters scored each rubric (low and high scores would be the same value when the initial scores were the same). In total 447 segments were analyzed for the ICC.

Table 1: Reliability Statistics (Absolute Agreement)

Academic Support Instruction Rubrics	Intraclass Correlation (Average Measures)	Social Support Instruction Rubrics	Intraclass Correlation (Average Measures)
AS1: Sense-making & reasoning	.618	SS1: Whole-class discourse	.809
AS2: Connections: representations & strategies	.602	SS2: Small-group discourse	.737
AS3: Pressing students to explain	.661	SS3: Status-raising / positioning students	.675
AS4: Context of tasks	.879	SS4: Motivational discourse	.581
AS5: Mathematical correctness	.378	SS5: Enthusiasm about mathematics	.524
AS6: Mathematical language precision	.548	SS6: Attention to students’ lives	.476
AS7: Feedback	.530	SS7: Accountability & high expectations	.505
AS8: Agency and autonomy	.620		

Note: N = 447

The initial rubric scores held an ICC of between .378 and .879, with an average of .610 (Table 1). Two rubrics (*mathematical correctness* and *attention to students' lives*) had poor reliability (.378 and .476, respectively). The remaining rubrics had moderate reliability of between 0.5 and 0.75 except for AS4 (*context of tasks*) and SS1 (*whole-class discourse*) which had good reliability (.879 and .809, respectively). Recall that ICCs were calculated based on initial ratings, prior to resolution meetings.

The relatively low ICC values for the majority of the scales indicated a need for some process of resolution. This led to three important questions: 1) Was a resolution meeting necessary, wherein the source of discrepancy and its nature are discussed, when the mean of the raters' initial scores could suffice?; 2) In instances when raters' initial scores differed, what was the magnitude of the discrepancy?; and 3) Regarding the nature of disagreements, did disagreements reflect attention to different evidence or disagreements about interpreting the rubric?

To address the first question, the differences between the mean of raters' initial scores and the final resolution score were analyzed for instances when initial agreement was not achieved. While individual differences would be expected here, in the aggregate such differences would balance out if the resolution meetings held no consistent sway on the resolved score--i.e., if the discrepancy were random error. The second question was answered by looking at the magnitude of any initial disagreements - describing whether these disagreements were mostly of a single rubric point or whether they represented greater disagreement among raters.

To answer the third question, raters were asked to describe the nature of any initial disagreements with observations resolved in the Spring of 2020 onward. This resulted in such data for 70 different segments, or 1,050 resolved scores spread across the 15 rubrics. For each rubric in every segment where there was a disagreement, raters identified whether this resulted from individuals observing the same evidence and rating it differently or whether different raters identified different aspects of the same phenomenon resulting in different initial scores.

Results

The most frequent outcome for resolutions was an increase of 0.5 relative to the mean of the initial scores for each rubric. The exception to this was *mathematical correctness*, which most frequently dropped 0.5 points and *accountability and high expectations* which most frequently resolved to the initial mean. The most common difference in initial ratings was 1, which held true for every rubric in the observation tool. Together these results show that, when disagreements occurred, they tended to be minor and the resolution discussions tended to result in agreement on the higher score (e.g., initial scores of 1 and 2, with a mean of 1.5, would be expected to resolve to a 2).

Within the 1,050 resolution scores analyzed to understand the nature of such disagreements, 405 initial disagreements occurred. Among these, 60 (15%) occurred when raters observed the same evidence but still initially disagreed on their rating and 346 (85%) occurred when raters observed different evidence which influenced their initial scores. The fewest disagreements arose for *whole-class discourse* (8, with 0 for same evidence and 8 for different evidence) and the most arose for *connections with representations and strategies* (40, with 4 for same evidence and 36 for different evidence).

Discussion

Our process of establishing validity suggests that the SMiLES project's observation tool has potential for measuring mathematics instructional practices that are potentially engaging for secondary students. The rubrics align with prior research about engaging mathematics instruction, as suggested from both internal and external conceptual validity investigations. This tool offers a set of rubrics that differs from existing observational tools designed to investigate high quality mathematics instruction for supporting students' learning.

Our external validity study afforded an opportunity to reflect on mathematics teaching for equity in relation to mathematics engagement. Although access to high quality mathematics instruction is important for equitable teaching and learning, it is a limited conception of equity. We revised our observation tool to capture how teaching could potentially support development of students' identities to address another dimension of equity (Gutiérrez, 2002).

Regarding reliability, the moderate to good ICCs for all but two of the rubrics showed general agreement of raters prior to resolution meetings, but not to an extent that would justify removing the resolution meetings from the observation analysis process. When disagreements did arise, they were typically minor but still afforded valuable insight when resolving scores. In some cases (~15%) the coders had conceptual differences in understanding mutually observed evidence, but in most cases (~85%) one coder had captured additional evidence which strengthened the justification of the final, resolved score.

Such results support the original intention of the resolution meetings as a way to ensure that various manifestations of these instructional supports are actually captured from the data. The data was not just "double coded" and averaged by the research team, but rather every single rubric score was discussed and agreed upon. Conceptual differences were thus addressed continually as they arose in the data, and coders had opportunities to gauge the *sum* of their evidence before committing to a rating. Since resolved ratings trended higher than the mean of the initial ratings, this could indicate that these meetings uncovered more evidence of potentially engaging mathematical instructional practices than otherwise would have been revealed.

Reliability training and double coding are valuable tools for qualitative research, but they do not transform qualitative analysis into an automated endeavor. The SMiLES project's resolution process identified one way in which the human capital of a research team can be utilized to strengthen analysis and more reliably capture relevant findings. Through this approach, disagreements are not a source of alarm but rather an opportunity to strengthen the foundation of the work itself. Initial ratings are not immutable but rather subject to interpretation and revision. Through this process, the complex nature of these engaging mathematical instructional practices – and, in turn, the work of these educators endeavoring to make them a reality – is better recognized.

The SMiLES project's observation tool demonstrates promise for investigating the presence and quality of potentially engaging instructional practices. The process we used when enacting the process of double coding provided a powerful approach for assessing instruction thoroughly. Events in a classroom are complex, and our research team found it helpful to have more than one coder noticing events that could be relevant. With this tool and this analytic process, perhaps the field can go further to understand how mathematics teaching can engage secondary students.

Acknowledgments

This research was supported by the National Science Foundation (EHR-Core) under Grant No. 1661180. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Anderson, A., Hamilton, R. J., & Hattie, J. (2004). Classroom climate and motivated behavior in secondary schools. *Learning Environments Research, 7*(3), 211–225.

Bostic, J. D., Matney, G. T., & Sondergeld, T. A. (2019). A validation process for observation protocols: Using the Revised SMPs Look-for Protocol as a lens on teachers' promotion of the standards. *Investigations in Mathematics Learning*, *11*(1), 69-82.

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

Chouinard, R., & Roy, N. (2008). Changes in high-school students' competence beliefs, utility value and achievement goals in mathematics. *British Journal of Educational Psychology, 78*, 31–50.

Collie, R. J., Martin, A. J., Bobis, J., Way, J., & Anderson, J. (2019). How students switch on and switch off in mathematics: exploring patterns and predictors of (dis) engagement across middle school and high school. *Educational Psychology*, *39*(4), 489-509.

Cohen, E. G., & Lotan, R. A. (1995). Producing equal-status interaction in the heterogeneous classroom. *American Educational Research Journal, 32*(1), 99–120.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, *20*(4), 399-483.

Fuentes, S. Q. (2018). Fostering Small-Group Discourse Student-to-Student Discourse. *Australian Mathematics Teacher*, *74*(2), 21-30.

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. *Educational Psychologist*, *50*(1), 14-30.

Gresalfi, M., Martin, T., Hand, V., & Greeno, J. (2009). Constructing competence: An analysis of student participation in the activity systems of mathematics classrooms. *Educational Studies in Mathematics*, *70*(1), 49-70.

Gutiérrez, R. (2002). Enabling the practice of mathematics teachers in context: Toward a new equity research agenda. *Mathematical Thinking and Learning, 4*(2–3), 145–187.

Hayashi Jr, P., Abib, G., & Hoppen, N. (2019). Validity in qualitative research: A processual approach. *The Qualitative Report*, *24*(1), 98-112.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horn, I. S. (2017). *Motivated: Designing math classrooms where students want to join in.* Portsmouth, NH: Heinemann.

Jansen, A., Middleton, J., Wiezel, A., Cullicott, C., Zhang, X., Tarr, G., & Curtis, K. (2019). Secondary mathematics teachers' efforts to engage students through academic and social support. In Otten, S., Candela, A. G., de Araujo, Z., Haines, C., & Munter, C. (Eds.). *Proceedings of the forty-first annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pgs. 1434-1443). St Louis, MO: University of Missouri.

Kazemi, E. & Stipek, D. (2001). Promoting conceptual understanding in four upper elementary mathematics classrooms. *Elementary School Journal, 102,* 59–80.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, *13*(2), 129-164.

Kosko, K. W. (2016). Primary teachers' choice of probing questions: Effects of MKT and supporting student autonomy. *International Electronic Journal of Mathematics Education, 11*(4), 991-1012.

Lubienski, S. T. (2002). Research, reform, and equity in US mathematics education. *Mathematical Thinking and Learning*, *4*(2-3), 103-125.

Middleton, J., Jansen, A., & Goldin, G. (2017). The complexities of mathematical engagement: Motivation, affect, and social interactions. In J. Cai (Ed.) *First Compendium for Research in Mathematics Education* (chapter 25, p. 667-699), Reston, VA: NCTM.

Nathan, M. J., & Knuth, E. J. (2003). A study of whole classroom mathematical discourse and teacher change. *Cognition and Instruction*, *21*(2), 175-207

Portney, L.G.,& Watkins, M.P. (2000). *Foundations of Clinical Research: Applications to Practice (2nd edition)*. Upper Saddle River, NJ: Prentice Hall.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, *102*(6), 245-253.

Schiefele, U., & Csikszentmihalyi, M. (1995). Motivation and ability as factors in mathematics experience and achievement. *Journal for Research in Mathematics Education*, *26*(2), 163-181.

Shernoff, D. J., Kelly, S., Tonks, S. M., Anderson, B., Cavanagh, R. F., Sinha, S., & Abdi, B. (2016). Student engagement as a function of environmental complexity in high school classrooms. *Learning and Instruction, 43*, 52–60. https://doi.org/10.1016/j.learninstruc.2015.12.003

Shernoff, D. J., Ruzek, E. A., & Sinha, S. (2017). The influence of the high school classroom environment on learning as mediated by student engagement. *School Psychology International*, *38*(2), 201-218.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, *33*(2), 455-488.

Stipek, D., Salmon, J. M., Givvin, K. B., Kazemi, E., Saxe, G., & MacGyvers, V. L. (1998). The value (and convergence) of practices suggested by motivation research and promoted by mathematics education reformers. *Journal for Research in Mathematics Education*, 465-488.

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, *94*(1), 88-106.

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, 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.

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The Math and Science Engagement Scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16-26.

Wigfield, A., Eccles, J. S., Mac Iver, D., Reuman, D. A., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology*, *27*(4), 552-565.

Yamauchi, L. A., Wyatt, T. R., & Carroll, J. H. (2005). Enacting the five standards for effective pedagogy in a culturally relevant high school program. In *Learning in Cultural Context* (pp. 227-245). Springer, Boston, MA.