Development and Initial Analysis of the Developing Mathematical Thinking Observation Instrument

NCTM Research Presession April 2014

Michele B. Carney
Boise State University

Jonathan L. Brendefur
Boise State University

Keith Thiede
Boise State University

Gwyneth Hughes
Boise State University

Brady Webb
Boise State University

Angie Crawford
Boise State University

Send correspondence to: Michele Carney, PhD
Boise State University
1910 University Drive
Boise, ID 83725-1745
michelecarney@boisestate.edu
Abstract

This paper describes the design and development process for the Development Mathematical Thinking observation instrument. It also includes initial psychometric analyses that were conducted. Our initial analysis indicated good to high levels of inter-rater reliability and internal consistency when the instrument was used by full-time project staff. However, the level of inter-rater reliability decreased greatly for our second set of analyses. This can likely be attributed to two potential factors; (1) use of raters who were familiar with the instrument framework but not full-time project staff, and/or (2) use of videos of lessons from teachers not actively involved in a professional development activity similar to our framework. Next steps in instrument development are addressed.
Introduction

Recent policy initiatives, such as the wide-spread adoption of the Common Core State Standards (CCSS) and associated accountability measures (PARCC and SBAC), have increased the pressure for teachers, schools and districts to examine classroom instructional practice. Practitioners and researchers need to determine what aspects of instruction influence students’ learning of the type and depth of mathematics called for by these new standards and assessments (Schmidt, 2012). These organizations have made it clear that to be considered ‘proficient’ in mathematics, students should be able to reason and communicate with mathematics, make connections within and outside of mathematics, and to persevere in problem solving (Burkhardt, Schoenfeld, Abedi, Hess, & Thurlow, 2012; NGA & CCSSO, 2011). For decades, the mathematics education research community has been calling for mathematics instruction that supports students in developing these practices; now that type of instruction is also supported and assessed through federal and state policy. However, large-scale classroom observations indicate that mathematics instructional practices may not widely match what is needed for students to achieve this new level of mathematics proficiency (Kane & Staiger, 2012; Stigler & Hiebert, 1999).

As districts work to implement the CCSS and address the requirements of the new assessments, it is essential they have tools that provide feedback on classroom instruction and support teachers’ reflection and growth (Measures of Effective Teaching Project, 2010a). In addition, with the plethora of professional development opportunities available around the new standards, we need to determine what types and aspects of professional development are significantly influencing practice (Hill, Beisiegel, & Jacob, 2013). We need multiple measures
available that evaluate classroom instruction – and specifically mathematics instruction – from various perspectives in order to provide useful feedback to professional developers, districts and teachers about classroom practice.

Our group has engaged in several mathematics teacher professional development projects over the past 10 years. Through our extensive work with local school districts (Brendefur, 2007, 2011) and a 3-credit, mandated state-wide professional development course (Carney, Brendefur, Thiede, Hughes, & Sutton, In Review), we have developed a framework for mathematics instructional practice focusing on developing students’ mathematical understanding. Based on our desire to quantify changes in teachers’ practice as a result of our professional development, we developed a classroom observation instrument to evaluate our professional development around this framework. The purpose of this paper is to present the theoretical framework, development process, and initial analyses of an observation measure designed to examine mathematics instruction and to provide the next steps in its development.

**Observation Instruments Designed to Measure Classroom Instruction**

Several observation-based measures of instructional practice exist in the literature. A sample of three commonly referenced measures is available in Appendix A. These surveys represent a wide range of purposes and perspectives on instructional practice.

Charlotte Danielson’s Framework for Teaching (FFT) identifies and measures the components of a teacher’s profession that promote improved student learning (Danielson, 2007). It is grounded in a constructivist perspective of learning and instruction. The complexity of teaching is analyzed through four domains each with multiple sub-constructs; (1) planning and preparation, (2) classroom environment, (3) instruction, and (4) professional
responsibilities. It is designed as both an evaluative tool and instrument for fostering professional conversations. It is utilized across multiple content areas, settings and activities. Scores on the Frameworks observation instrument have been linked to student achievement (Measures of Effective Teaching Project, 2012). To achieve reasonable levels of validity and reliability, one must use the Framework with fidelity and undergo comprehensive training which is a relatively lengthy and costly process.

The Reformed Teaching Observation Protocol (RTOP) was developed to measure the use of ‘reform practices’, particularly inquiry, in mathematics and science classrooms (Sawada et al., 2002). The instrument has 25 items across three domains: lesson design and implementation, content, and classroom culture. The instrument is used during real-time observations and includes space for documenting classroom events. Sawada et al. (2002) provide instrument reliability and predictive validity evidence related to student learning. The instrument is utilized in both research settings and as a professional development tool with teachers. The instrument, training guide, and reference manual are all available online.

The Mathematical Quality of Instruction (MQI) instrument is used to measure the character of a teacher’s classroom interactions that shape mathematics instruction and student learning (Learning Mathematics For Teaching, 2011; Measures of Effective Teaching Project, 2010b). The instrument framework consists of seven major constructs and their corresponding scales; (1) richness of the mathematics, (2) development of the mathematics, (3) responding to students, (4) connecting classroom practice to mathematics, (5) language, (6) equity, and (7) presence of unmitigated mathematical errors. The MQI framework and development process
focused on a pedagogically neutral approach to observation of mathematics practices and focuses on identifying constructs that form separate variables in order to explore whether and how each construct affects student learning. The MQI instrument implementation consists of an extremely labor intensive process of training and implementation.

DMT Observation Instrument Theoretical Framework

The Developing Mathematical Thinking (DMT) observation instrument was developed over a period of four years. The instrument is designed to measure five dimensions of practice related to mathematics instructional quality. Specifically, it measures dimensions related to engaging students in meaningful, intellectual activities that provide authentic understanding of mathematics and builds connections both within and outside the classroom. The underlying five-part framework is based on social-constructivist theories of learning and presumes students must construct mathematical knowledge through guided instruction in order to use mathematics with meaning. Unlike a discovery learning or radical constructivist approach, the role of the teacher is one of a facilitator of knowledge who must provide students with rich, meaningful tasks, and press their level of sophistication and abstraction through a continual focus on the underlying mathematical structure. Unlike the MQI observation instrument we embed very specific pedagogical ideas within the observation instrument descriptors. The specific dimensions within the framework are:

Taking students' ideas seriously (TSIS), involves valuing and building upon students’ intuitive understanding of mathematical concepts (Carpenter & Lehrer, 1999; Hiebert, 1997; Romberg & Kaput, 1999). For example, when students solve an unfamiliar yet meaningful math problem, they draw on their prior knowledge and experience. Their solution strategies and
notations may seem inefficient or informal to an observer, but by eliciting and valuing students’ initial solution strategies, teachers can connect student thinking to more efficient and abstract methods (Freudenthal, 1973, 1991; Gravemeijer & van Galen, 2003; Treffers, 1987).

**Encouraging multiple strategies and models (EMMS)**, involves developing students’ understanding of various models and approaches to solving problems (Dolk & Fosnot, 2006; National Council of Teachers of Mathematics, 2000; Romberg & Kaput, 1999). When students generate, evaluate, and utilize different mathematical strategies and models, they recognize there are many ways to solve problems and represent solutions. In addition, different strategies and models highlight different aspects of the mathematics and thus examining the same problem through different lenses deepens students’ overall understanding of the topic.

**Pressing students conceptually (PSC)**, focuses on building connections between mathematical strategies and models and progressively formalizes those ideas and methods for solving problems (Carpenter & Lehrer, 1999; Forman, 2003; NRC, 2001; Siegler, 2003). For example, once students have had the chance to work on their own solution methods, teachers press them to connect and compare between methods, generalize to new situations, and relate to formal mathematical terms and conventions. It is through this process of connection and generalization that moves students from their own informal methods to more formal and efficient strategies (Carpenter & Lehrer, 1999; Gravemeijer & van Galen, 2003).

**Addressing misconceptions (AM)**, involves using students’ mistakes and misconceptions as valuable tools to build mathematical understanding (Borasi, 1987, 1994; Bray, 2013; Gooding & Stacey, 1992). Making mistakes and learning from them is an integral part of doing mathematics at any level. But mistakes often recur even after teachers demonstrate a correct
procedure because they stem from deeper mathematical misconceptions. By being aware of why and how misconceptions develop and taking the time to address misconceptions through models and discussion, teachers can move students to a deeper level of understanding that precludes such mistakes. Additionally, mistakes can be opportunities for students to engage in justification, evaluation, and inquiry (Borasi, 1987).

**Focusing on the structure of the mathematics (FSM),** involves facilitating students' understanding of fundamental, or structural, mathematical concepts (e.g., decomposing and composing, units and unitizing, equivalence, etc.). Focusing on these structures allows students to build understanding of and establish connections between these fundamental concepts and the particular topic being studied (National Council of Teachers of Mathematics, 2000; NGA & CCSSO, 2011). Many teachers and their students see mathematics as a series of procedures and definitions that build in complexity throughout the K-12 curriculum. But certain fundamental ideas or “structural components” appear continually throughout mathematics, whether one is looking at 2nd or 11th grade. Understanding these structural components can help teachers tie different concepts together both within and across grade levels, rather than teaching topics in isolation. When instruction does not focus on the structure of mathematics, students often rely on memorized tricks or formulas and have difficulty solving complex problems or applying mathematics to new situations.

**Methods**

**Instrument Development**

The domains of the DMT framework provided the five areas of scale development. We were unsure whether the domains would emerge as separate variables that can be measured
independently or whether they all measure an overall construct related the DMT framework.

For the purposes of instrument development, each DMT domain is conceptualized as a separate construct that is measured with a separate scale. Each scale then consists of four sub-constructs with corresponding sub-scales. A particular sub-scale for a sub-construct has an overall descriptor followed by specific item descriptors for each score ranging from 1-5 (see Table 1 for an example). The scales and sub-scales have gone through an iterative, multi-year process of administration and revision.

Table 1: Item descriptors for sub construct “Classroom activities are focused on, and adapted to, the responses and experiences of the students” within the “Taking Students Ideas Seriously” domain.

| A. Classroom activities are focused on, and adapted to, the responses and experiences of the students |
|---|---|
| | Description |
| 1 | Classroom activities are guided by the textbook or other resource. |
| 2 | Classroom activities take into account some of the experiences of students, but do so superficially. The scope and sequence of the unit are still primarily addressed using the textbook. |
| 3 | Classroom activities take into account some of the experiences of students. Although the text provides the majority of the scope, sequence, and tasks, the teacher makes an attempt to modify activities to meet the needs and experiences of students. |
| 4 | The teacher is able to modify classroom activities as needed based on the responses and experiences of the students. |
| 5 | The teacher is able to clearly demonstrate that classroom activities were selected based on the experiences of students, and that additional supports and extensions have been generated prior to the lesson based on this knowledge. |

Experiments

Our initial analyses of the DMT observation instrument involved an iterative process.

Two cycles of analysis occurred that utilized two different data sources. For the sake of clarity, we describe the data sources and results from the initial analysis as ‘experiment 1’ followed by the data sources and results from the second analysis as ‘experiment 2’.

Experiment 1

The data for experiment 1 came from a three year state-level Math and Science Partnership project that focuses on extending teachers' knowledge of mathematics, their understanding of how children learn mathematics, and their ability to teach mathematics in a
more effective manner. The 65 grant participants received a 45 hour mathematics professional development institute each summer (for three summers) and ongoing support throughout the school year. Teacher observations were conducted in the fall and spring of each year by two project researchers. The initial analysis involved examining the inter-rater reliability and internal consistency of the data from the observations.

Data from the fall 2010 observations were utilized. A single measure ICC was used to calculate the inter-rater reliability (see Table 2) based on the recommendation of Hallgren (2012).

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sample Size</th>
<th>Number of Raters</th>
<th>Mean[SD]</th>
<th>Single Measure ICC</th>
<th>Average Measure ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taking Students’ Ideas Seriously</td>
<td>5</td>
<td>2</td>
<td>3.05(.57)</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Pressing Students Conceptually</td>
<td>5</td>
<td>2</td>
<td>2.73(.55)</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>Encouraging Multiple Models and Strategies</td>
<td>5</td>
<td>2</td>
<td>2.78(.51)</td>
<td>0.65</td>
<td>0.79</td>
</tr>
<tr>
<td>Addressing Misconceptions</td>
<td>5</td>
<td>2</td>
<td>2.55(.55)</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>Focusing on the Structure of the Mathematics</td>
<td>5</td>
<td>2</td>
<td>2.65(.36)</td>
<td>0.56</td>
<td>0.71</td>
</tr>
<tr>
<td>DMT Instrument Total</td>
<td>5</td>
<td>2</td>
<td>2.75(.45)</td>
<td>0.91</td>
<td>0.95</td>
</tr>
</tbody>
</table>

It is important to note that there is wide variability in how inter-rater reliability is calculated and reported in the educational research literature. While the inter-rater observation sample for experiment one was small, the ICC’s were good to excellent across the two observers for five of the six constructs. Cronbach’s alpha was used to calculate the internal consistency (see Table 3) (Cicchetti, 1994). The alphas were good to excellent across all domains for the entire instrument.

Table 3: Cronbach’s alpha for internal reliability of experiment 2.
While this analysis provided initial evidence of good inter-rater reliability and internal consistency, we had several additional items that could not be answered with this data set.

1. How generalizable is the inter-rater reliability? We wanted to determine if observers familiar with the DMT framework but not a full-time member of the research project could provide reliable observation scores.

2. Would the anticipated structure of the observation instrument around the DMT framework emerge from a factor analysis process? Experiment 1 data set contained 60 participants with multiple observations. In order to conduct a factor analysis we needed a minimum of another 40 participants.

3. What is the relationship between the DMT instrument observation scores (and sub-scores) to other student and teacher level variables of interest?

In order to address these questions, we applied for access to the Measures of Effective Teaching project database to utilize their videos of classroom mathematics instruction and the data from various measures collected at both the teacher and student level.

**Experiment 2**

The data source for experiment 2 came from the Measures of Effective Teaching project (Measures of Effective Teaching Project, 2010a, 2012, 2013). The purpose of the MET study is
to identify and develop effective teaching. The data gathered through the project is extensive and an application must be submitted to receive access to the data. We specifically applied to utilize the videos of classroom practice for 4th – 8th grade teachers’ mathematics lessons and related variables of interest at both the teacher and student level. For experiment 2 we utilized videos from year 1 of the MET project. We selected videos for observation based on a stratified sampling approach; low (1.0 to <1.5), medium (1.5 to <2.0), and high (2 to <2.5) based on their average MQI teacher observation scores. The MQI scores range from an available metric of 1 to 3, however, the range of scores did not extend beyond a 2.5 average and a very limited number of lessons were in the high range. Given the nature of our instrument and the fact it is designed to measure changes in practice as a result of intensive professional development, we felt a stratified approach was necessary to provide a range across the DMT instrument scale levels.

We selected two videos from the low and medium stratification levels to conduct our initial observer training. The trainer was a DMT project staff member with extensive background in the development and use of the instrument who was considered the ‘expert-rater’. The observers were a faculty and DMT project member and 3 graduate students with reasonable familiarity with the DMT framework through coursework and other various activities. The initial observer training involved thoroughly reading the DMT observation instrument domains and descriptors, followed by watching and coding the two videos. This was followed by a meeting where the compiled scores were provided to everyone in conjunction with the scores from the ‘expert-rater’. Each item score was thoroughly discussed in relation to the classroom video and clarification on descriptor meanings occurred throughout the process.
Following this meeting, everyone viewed and rated another 3 videos; 1 medium and 2 high from the stratified sample. This sampling was based on the low range of scores for the first set of observations. A similar process of discussion and analysis followed for this set of videos and was followed by a third cycle involving 2 videos. Table 4 provides the ICC single measure scores resulting from the 2nd and 3rd rounds of video coding across the 4 observers.

Table 4: ICC statistics for inter-rater reliability for experiment 2.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Sample Size</th>
<th>Number of Raters</th>
<th>Mean(SD)</th>
<th>Single Measure ICC</th>
<th>Average Measure ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taking Students’ Ideas Seriously</td>
<td>5</td>
<td>4</td>
<td>1.58(58)</td>
<td>0.57</td>
<td>0.84</td>
</tr>
<tr>
<td>Pressing Students Conceptually</td>
<td>5</td>
<td>4</td>
<td>1.40(54)</td>
<td>0.19</td>
<td>0.48</td>
</tr>
<tr>
<td>Encouraging Multiple Models and Strategies</td>
<td>5</td>
<td>4</td>
<td>1.27(35)</td>
<td>0.18</td>
<td>0.47</td>
</tr>
<tr>
<td>Addressing Misconceptions</td>
<td>5</td>
<td>4</td>
<td>1.73(72)</td>
<td>0.13</td>
<td>0.38</td>
</tr>
<tr>
<td>Focusing on the Structure of the Mathematics</td>
<td>5</td>
<td>4</td>
<td>1.46(51)</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>DMT Instrument Total</td>
<td>5</td>
<td>4</td>
<td>1.49(48)</td>
<td>0.39</td>
<td>0.72</td>
</tr>
</tbody>
</table>

As is evident from the extremely low ICC scores, we determined that the amount of inter-rater reliability was too extreme and further work needed to be conducted before continuing with the current process.

Conclusions

The purpose of this paper is to present the theoretical framework and initial psychometric evaluation of a mathematics instructional practice observation document. Initial analyses indicate promising findings in terms of inter-rater reliability and internal consistency when conducted by DMT project staff members intimately familiar with the development and implementation of the instrument. However, either the use of videos outside the professional development project which may have reduced the range of scores (and therein the level of inter-rater reliability that could be achieved) or the use of observers beyond DMT project staff or some combination of these factors created significant issues with the inter-rater reliability.
Due to these issues we have identified the following ‘next steps’ as (1) refine and further develop definitions and descriptions based on feedback from ‘non-research project’ coders, and (2) create a set of master coded videos for training purposes.

With the continuing focus on improving mathematics education and the large-scale implementation of assessments designed to measure students’ mathematics understanding at a deeper, and more conceptual level, it is vital that we have measures that can be used to inform both our understanding of teachers’ practice and the effect of mathematics professional development. We must have a variety of measures that focus on different aspects of mathematics practice in order to identify and further develop our understanding of effective teaching. While the results of our present study do not currently support the use of our observation instrument for a broader audience, we do intend to continue this work in the future.
References


<table>
<thead>
<tr>
<th>Instrument</th>
<th>Purpose</th>
<th>Domains (Number of Items)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Framework for Teaching (FFT)</strong></td>
<td>Identifies and measures the components of a teacher’s profession that promote improved student learning across multiple academic subjects</td>
<td>Planning and preparation (6) Defines how a teacher designs instruction and organizes the content. All elements of the instructional design – learning activities, materials, assessments, and strategies – should be appropriate to both the content and the learners.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classroom environment (5) Consists of the non-instructional interactions that occur in the classroom. Activities and tasks establish a respectful classroom environment and a culture for learning.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Instruction (5) Consists of the components that actually engage students in the content. These components represent distinct elements of instruction. Students are engaged in meaningful work that is important to students as well as teachers.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Professional responsibilities (6) Encompasses the teacher’s role outside the classroom. These roles include professional responsibilities such as self-reflection and professional growth, in addition to contributions made to the schools, districts, and to the profession as a whole.</td>
</tr>
<tr>
<td><strong>Reformed Teaching Observation Protocol (RTOP)</strong></td>
<td>Measure “reform” practices in mathematics and science instruction</td>
<td>Lesson design and implementation (5) Focuses on structure and design of reform instruction; respect for student ideas; community of inquirers; exploration before explication.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Content (10) Separates pedagogical content knowledge into propositional knowledge and procedural knowledge.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Classroom culture (10) Communication is diverse and decentralized; student-teacher relationships are egalitarian.</td>
</tr>
<tr>
<td><strong>Mathematical Quality of Instruction (MQI)</strong></td>
<td>Measures the mathematical quality of mathematics instruction, defined as the nature of the mathematical content provided to students during instruction (not measuring pedagogical approach)</td>
<td>Richness of the mathematics development of the mathematics (7) Students should be exposed to multiple representations that are coordinated and linked to each another; connections are made between multiple models; explanations; justifications; and explicit talk about mathematical language, reasoning, and practices.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Responding to students (2) Teachers’ responsiveness to student ideas and misconceptions; interprets student work and expands upon student errors.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Connecting classroom practice to mathematics (3) The work students engage in links to a mathematical idea or procedure; instructional time is spent on mathematics, and the amount of time teachers spend in mathematically meaningful lessons.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language (4) Conventional notation and technical language are present and appropriately used in instruction; whether notation or terms are simply used during instruction or whether there is explicit talk about their meaning.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equity (3-4) Teachers make sure that all students have access to the mathematics in class and are aware of differences among students’ mathematical foundations; instruction is explicit.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Presence of unmitigated mathematical errors (8) Computational errors or significant omissions as they occurred, such as providing an inaccurate presentation of content or performing a procedure incorrectly.</td>
</tr>
</tbody>
</table>