

Section 3: Task Development Related to the Validity and Reliability of the Instrument

Establishing Validity and Reliability of the PAM

Validity. The degree to which an assessment measures the characteristics it is designed to detect is never fully confirmed because establishing validity is a process that evolves while using an instrument over time (Gable, 1986). Three major types of validity are at the core of understanding an instrument's accuracy: content validity, criterion-related validity, and construct validity. The content of an instrument is the first concern of any test developer. In this case, the PAM was developed to assess the content knowledge of future teachers and their competency in using instructional strategies (pedagogical skills) to teach mathematics. Specifically, the purpose of the instrument is to assess pre-service teachers' abilities to use appropriate language related to mathematics, to apply appropriate instructional methods and materials, and to diagnose mathematical problems of students while imparting accurate mathematical concepts using the content and teaching standards of the National Council of Teachers of Mathematics (NCTM) and the Connecticut's Common Core of Teaching (CCT) (State of Connecticut, State Board of Education, 1999). These and other standards are aligned with the PAM in Appendix G. Furthermore, the content of the task and scoring rubric was examined by all members from both Team 1 and Team 2. Feedback from three checkpoints (summer 2002, fall 2002, and spring 2003) in the instrument development process was used to make changes to the task and rubric (Refer to Table 2). These three periods were the times that the instrument underwent pilot testing. Not only did members of each Team provide feedback, the degree of clarity reflected in the responses of the students supplied necessary information for modifying the task.

The purpose of having WestConn faculty and local mathematics specialists on Team 2 was to tap them as a valuable resource for obtaining feedback to the task and rubric. They served

as content experts regarding mathematical content and instructional strategies. Plans to garner more objective feedback from content experts at other institutions are under way. To date, a proposed list of the following individuals will be sent the task and scoring procedures for review:

Dr. Bruce Shore, McGill University, Montréal, Canada (mathematics and cognitive psychology); Dr. Thomas Williams, West Virginia Wesleyan College (education); Dr. Jose Hamer, Western Connecticut State University (mathematics); Dr. Michael Hibbard, Assistant Superintendent, Ridgefield Public Schools (performance assessment), Mr Donald Trocolla, Danbury Public Schools, Director of Mathematics.

In order to discern the evolution of the PAM’s development during the three pilot periods, the following questions and responses reveal the types of discussions and decisions that were made to alter the content and format of the instrument. Refer to Tables 4, 5, and 6.

Table 4

Questions and Responses Regarding the First Pilot Test of the PAM

Question	Response
1. Did students understand how to complete the task?	Students in the pre-student teaching semester were asked by the course instructor, “How do you feel about the assignment?” Many responded that this was a good review of their knowledge, that this is an appropriate task for future teachers, and that they felt prepared to complete the project.
2. How does this task fit into a student’s view of what should be taught in the classroom and how it should be taught?	The task was placed into the framework of a lesson.
3. What lesson format will be used to structure the task?	The lesson format from the Department of Education and Educational Psychology will supply the structure.

Question	Response
4. How should candidates express dialogue when they do not have an indication of how 4 th grade students will respond to the lesson?	Candidates are instructed to record “ideal” responses in order to understand how they think a student would or should respond.
5. What happens when a student writes too much? How can legibility of responses be adequately handled by the WestConn faculty who will score these responses?	Students are given a space limit.
6. Will student handwriting interfere with the recording of the student responses and the scoring of these responses?	The task was loaded into the lab computers and students typed their responses.
7. How should the faculty scoring the responses indicate their ideas about the quality of the response?	A coding/scoring area was added to every page of the response.
8. Does the scoring rubric directly relate to the task? Is it easy to use?	The scoring rubric had more levels of categories than would have been expected by the respondents. For example, the students would not have known to include all types of representations. Discussion began regarding how to simplify the rubric.
9. Based on the Team’s initial rubric, how are students to know how to respond to the task as completely as possible?	The mathematics professors all began to include information about the task in their courses. Professors in the education department began to incorporate the specific ideas related to the rubric in their revised course outlines and syllabi.

Table 5

Questions and Responses Regarding the Second Pilot Test of the PAM

Question	Response
1. Is this format too cumbersome? Are too many directions needed to satisfactorily complete the task?	The lesson format was reviewed to address its purpose in producing an adequate response. The lesson format was dropped for a trimmed format focusing on the content of a CMT or CAPT problem.
2. Is the Guided Practice section necessary?	It seems redundant with the information produced in the initial Script/Dialogue section. The lesson framework was abandoned since it was too complex and redundant.
3. Is the content of the task, introduction to division, an important enough topic to be included in this assessment?	The focus of the content was changed from a “typical” lesson topic to a CMT or CAPT topic. Since these released items already have scored student responses, students will be practicing a skill needed in their future classrooms.
4. Which CMT and CAPT items will be appropriate for the task?	Theoretically, any randomly selected CMT or CAPT released item should be useful for this task. The example student response representing a “2” will be used in the task since this level has adequate error (not too much, not too little) for a candidate to formulate a response.
5. Will this format easily transfer to a secondary task?	The underlying intention of the task has not changed, but the focus of a “typical” mathematics topic to a CMT/CAPT topic provides more focus for the transfer from an elementary to a secondary task.
6. Is this scoring rubric easy to use?	The NCTM topics of communication, connections, problem-solving, reasoning and proof, and representations remain the focus of the task. These terms need to be more closely related to the purpose of the task, the instructions for the task, and the anchor ideas for assessing the responses.

Question	Response
7. Does this task validly assess the ability to know and teach mathematical content?	More alignment needs to be completed between the purpose of the task, the rubric, and the scoring procedures.
8. Did students understand how to complete the task?	When asked to supply comments about the project, the students in the pre-student teaching semester responded that they felt prepared to complete the task.
9. Were students adequately prepared to complete the task?	Based on a qualitative analysis of responses, students in the Professional Semester (pre-student teaching) were able to respond more satisfactorily regarding the types of instructional strategies they would use in the classroom than were the mathematics students who had not completed any courses in the methods of teaching.
10. How will all raters understand the procedures asked of the task?	All raters were provided with an individualized review of how to use the rubric. A question and answer period was available and scoring was conducted.
11. Is there inter-rater reliability when using the scoring rubric?	Plans to improve the scoring rubric based on the improvements to the task were developed. Raters from Team 1 and Team 2 scored summer 2002 responses using the spring 2003 rubric. Inter-rater reliability was calculated. Refer to Table X.

Table 6

Questions and Responses Regarding the Third Pilot Test of the PAM

Question	Response
1. Did all raters understand how to use the rubric?	All raters were provided with an individualized review of how to use the rubric. A question and answer period was available and scoring was conducted.
2. How will inter-rater reliability be recorded when between scorer variance was so consistent that there was not enough variance to calculate the reliability?	Inter-rater agreement was employed. Refer to Table X and Table X for results.
3. Did students understand how to complete the task?	Due to time constraints, a formal assessment of student feedback was not completed. Students responded that felt adequately prepared to complete the assignment, when asked by their course instructor.
4. Does the rubric match the content of the task?	All raters took notes while using the task. This information will be reviewed.

Students completing the most recent version of the PAM were asked to respond to an evaluation of the task. Since PAM-related activities have not yet been fully integrated into mathematics and education courses, it was not surprising to find that students stated that they did not necessarily feel ready to complete the task. Nonetheless, 11 out of 14 students responded that this was an appropriate task for an education major. See Table 7 for complete results.

Table 7

Student Evaluation of the PAM Task in May 2003.

CANDIDATE FEEDBACK SURVEY RESULTS

Elementary Education (n=9) Secondary Math (n=7) No Item Checked (n=1)

1. Demographic Information:

No reported Here

2. For pre-student teaching elementary education candidates, this assessment was:

too difficult

**somewhat
challenging**

appropriate

too easy

1

2

3

4

(n=1)

(n=2)

(n=11)

Average of responses: 2.714

Three students provided no response for #2.

Average Rating (n=17)

Average of responses:

*Strongly
Agree*

Agree

Disagree

*Strongly
Disagree*

1

2

3

4

3. The assessment took an appropriate amount of time to complete.	2.294
4. I had sufficient learning opportunities in my program, prior to this assessment, in order to successfully complete it.	2.118
5. This assessment required me to demonstrate skills that are essential to my success as a teacher.	1.706
6. This assessment required me to analyze important reading comprehension skills.	2.235
7. This assessment required me to plan instruction based on student performance.	2
8. This assessment enabled me to evaluate and reflect upon basic pedagogical practice.	2
9. Assessments like this will be helpful to me in student teaching and relate to realistic expectations of me as a teacher.	1.706
10. This assessment required me to reflect on my practice.	1.882
11. My teacher preparation program (courses and field experiences) prepared me well for this assessment.	2.344

Plans for collecting criterion-related validity include correlations of PAM results with grades from required mathematics courses (MAT105 and MAT106 for elementary education students; GPA for secondary mathematics majors) and with student teaching evaluations of mathematics lessons. These data will be used to examine the hypothesis that pre-service teaching students who demonstrate capabilities in knowing mathematical content and in using instructional strategies in the classroom will provide acceptable responses on the PAM.

One way to assess construct-related validity will be to calculate the change in pre-service teaching students' responses on the PAM as they progress through the certification program. It is expected that students will have lower scores at the beginning of their coursework in mathematics education and that the PAM will detect increases in student performance prior to student teaching and be higher after student teaching.

Integrating the task into the Undergraduate Teacher Certification Programs at the Elementary and Secondary Levels. Since all certification candidates who are ready to do their student teaching should know the content they propose to teach and necessary strategies for teaching that content, all candidates should be able to provide adequate responses to the PAM at the "Fully Met" or "Met" level prior to student teaching. The content knowledge to teach mathematics is reviewed and reinforced in MAT105 and MAT106 for students focusing on their elementary education certification and in all courses in mathematics for students in the secondary education certification program. In the near future, these courses will include sample PAM tasks in order to familiarize students with the purpose of the task and the scoring expectations. It is planned that students will continue to encounter PAM tasks during their Professional Semester, when the methods courses are scheduled. The Professional Semester also provides elementary

education students with opportunities to have teacher mentors in the local public schools, as they practice teaching for a period of 10 days (see Appendix K for a copy of the Professional Semester Guidelines). Sample PAM tasks will be available for students to review and to compare to children's responses during the Professional Semester.

Near the end of the Professional Semester it is planned that all certification students will be administered the PAM. They would be given feedback and an opportunity to review any objectives not passed, followed by a re-administration of an alternate form of the PAM. For more information, also refer to Section Four: Connecting the Task to Unit Assessment, *The Process to institutionalize this task*.

Reliability. Inter-rater reliability is the degree of consistency of responses among raters. All raters were trained to use the PAM scoring rubric in two separate sessions, one for Team 1 members and 1 session for those on Team 2. During each of these sessions, all members read the task and the rubric noting comments and questions about the wording on the instruments and the scoring process. A list of common terms was made for each of the performance levels (Target Performance/Fully Met, Partially Met, Not Met, No Response) in order to assure that similar vocabulary would be used when scoring responses. These terms are listed at the end of the rubric. All team members were familiar with exemplary responses since each had recently provided what he or she thought was an exemplary response to the problem (see Appendix J for responses).

Reliability was first sought using the updated PAM rubric (spring 2003) to assess the agreement among raters for student responses to the PAM task developed in the summer of 2002. While not an ideal fit between the task and rubric, Part 3 and Part 4 of the PAM rubric were applied to the completed student responses of the earlier version of the PAM task. Intra-class

Correlation Coefficients were calculated using the Formula 1. Results for the first inter-rater reliability assessment for 6 pre-service teaching students' responses are contained in Table 8.

Formula 1

$$ICC(2) = \frac{MSB - MSW}{MSB}$$

ICC(2)=Intraclass Correlation Coefficient, reliability for mean ratings from *k* raters

MSB=mean square between

MSW=mean square within

(Guilford, 1954, p. 395; Klein & Kozlowski, 2000, p. 354-356)

Table 8
Interrater Reliability Coefficients for the First Version of the PAM Administered to Preservice Teachers in the Elementary Education Certification Program.

PAM Category	Reliability for Mean Ratings from <i>k</i> Raters ¹
Part 3: Content Pedagogy: Planning the Instructional Strategy	.98
Part 3: Content Pedagogy: Content Knowledge	.98
Part 4: Assessment: Analysis of Strengths and Weaknesses	.68
Part 4: Assessment: Instructional Strategy	.86
Part 4: Assessment: Feedback	.81

Note¹: Ten raters were included in these analyses.

Note²:Pre-service teaching students were not specifically instructed to identify strengths and weaknesses in the child's answer to this task. Thus, a lower reliability coefficient was found.

After revising the task in the spring of 2003, X pre-service teaching students in the elementary education certification program and 10 in the secondary education certification program were administered the revised PAM. Inter-rater reliability results from 12 raters are available in Table 9 for the elementary certification students and in Table 10 for the secondary certification students.

Table 9
Inter-rater Agreement for the Third Version of the PAM Administered to Pre-service Teachers in the Elementary Education Certification Program at WestConn.

PAM Category	Percent of Agreement Among k Raters* to Pass or Remediate
Part 1: Content Knowledge: Problem Solving, Mathematical Language and Skills	100%
Part 2: Content Knowledge: Content Concepts and Skills	98%
Part 3: Content Pedagogy: Planning the Instructional Strategy	89%
Part 3: Content Pedagogy: Content Knowledge	88%
Part 4: Assessment: Analysis of Strengths and Weaknesses	95%
Part 4: Assessment: Instructional Strategy	91%
Part 4: Assessment: Feedback	81%

*Note: 8 raters were included in these analyses. Raters had not yet had the opportunity to discuss their ratings to reach agreement. Additional discussions will be held to reach higher agreement and the task will be rescored.

Table 10
Inter-rater Agreement for the Third Version of the PAM Administered to Pre-service Teachers in the Secondary Education Certification Program at WestConn.

PAM Category	Percent of Agreement Among k Raters* to Pass or Remediate
Part 1: Content Knowledge: Problem Solving, Mathematical Language and Skills	86%
Part 2: Content Knowledge: Content Concepts and Skills	85%
Part 3: Content Pedagogy: Planning the Instructional Strategy	76%
Part 3: Content Pedagogy: Content Knowledge	81%
Part 4: Assessment: Analysis of Strengths and Weaknesses	90%
Part 4: Assessment: Instructional Strategy	76%
Part 4: Assessment: Feedback	79%

*Note: 8 raters were included in these analyses. Raters had not yet had the opportunity to discuss their ratings to reach agreement. Additional discussions will be held to reach higher agreement and the task will be rescored.

Future Scoring of the PAM. In order to train future raters of candidates' responses, an orientation to the PAM will be required. This process will include an overview of the purpose of the instrument, a question and answer period regarding the components of the PAM and its alignment with specific mathematical and educational standards. Finally, all prospective raters

will be asked to complete the PAM themselves and to score a range of sample responses. A discussion will be conducted concerning any responses that do not agree with other trainees and with the pre-scored response. Trainees will need to have 100% agreement on the issue of whether to score a Pass or to Remediate for each PAM category, prior to scoring new PAM responses.