

Section 3 Task Development

Section Overview:

The mathematics assessment consisting of 14 open-ended items and 10 multiple choice items was piloted with 39 incoming elementary teacher education candidates on December 11, 2002. The science assessment consisting of 23 open-ended items was piloted with 39 incoming elementary teacher education candidates on December 18, 2002. The same pool of candidates participated in both the mathematics and science assessments.

Prior to piloting the assessments with candidates, the items were reviewed by content experts (Mari Muri, CSDE mathematics consultant and Donna Leake, Ph.D., principal of Norton Elementary School, and Adrienne Kanach, Ed.D., project consultant and former science educator). See Appendix B: Documentation of Validity and Reliability (Project Members' Qualifications)

In addition, the science items were reviewed in a "pre-pilot" by five elementary teachers at Norton ES. After taking the test, these teachers completed a questionnaire providing feedback on each item's clarity and relevancy for elementary teachers' knowledge base. In addition, teachers provided general feedback about the length of the assessment and its validity as a fair assessment of elementary teachers' content knowledge. Similarly, the mathematics assessment items were "pre-piloted" by 15 Norton ES teachers as. These teachers provided feedback on the clarity of the items (See Section 2.VII Feedback Summaries).

Assessment items were also reviewed by TBA Consulting (see Section 2 for the TBA report and our response). In developing the final item pool, we considered the recommendations of consultants as well as item analyses of the December pilot.

Finally, candidates who participated in the December pilot will be asked to complete a brief questionnaire after they complete their program in the fall. The questionnaire will provide additional information about our strategy courses as well as the validity of the assessment and the degree to which they perceived the assessments a fair test of their mathematics and science content knowledge (See Appendix B).

Construct Validity

The assessments have been developed to assess preservice teachers' conceptual knowledge in mathematics and science concepts. In designing our assessments, we had to consider both construct validity as well as issues of fairness, accuracy, internal consistency, and alignment with state and national standards for elementary teacher candidate knowledge. Construct validity was a primary concern because our project's success rested on designing items and

scoring protocols that would sufficiently discriminate candidates' conceptual knowledge from factual knowledge.

We have defined **conceptual knowledge** as candidates' ability to explain how they reason about scientific and mathematical concepts. That is, they are able to provide mathematically accurate explanations of their mathematical solutions and to provide complete responses that correspond to scientifically accepted explanations for the phenomena described in science items (Stoddard, Connell, Stofflett, & Peck, 1993). In mathematics, for example, candidates were required to explain how they reasoned mathematically, solved problems, and identified patterns. In science, candidates were required to explain how one or more underlying scientific principles informed their understanding of scientific concepts or to use one or more components of the scientific inquiry process to reason scientifically.

Conceptual knowledge is important for teacher candidates to have if they are to encourage students to see central ideas in the disciplines and understand how these ideas connect to other disciplines and to real-life situations. Kennedy (1997) distinguishes the quantity (content) from the quality (conceptual understanding) of teacher knowledge. Conceptual understanding includes elaborated knowledge and reasoning ability as well as the ability to see the "big idea" or central ideas of the discipline.

Consistent with this view, our project focuses on the assessment of the quality of teacher knowledge, not the quantity. Our project builds on previous research conducted by Stoddard, Connell, Stofflett, and Peck (1993) in which researchers assessed the content knowledge of preservice teachers in mathematics and science. Three categories of mathematical content knowledge were identified and candidates were assessed as having (1) neither procedural or conceptual understanding, (2) procedural but not conceptual, and (3) procedural and conceptual understanding. Those with only procedural understanding arrive at the "correct" answer using algorithms or recall but are not able to explain their reasoning. Similarly, science responses were classified into three categories, (1) naïve conceptions, (2) scientifically naïve conceptions, and (3) scientific understanding. Those with naïve conceptions had no understanding, and those with scientifically naïve conceptions provided partially correct explanations, such as providing reasonably accurate explanation but using scientific terminology incorrectly. Those with scientific understanding were able to provide completely accurate scientific explanations.

Consistent with this previous research, our rubrics have been designed to identify similar categories of knowledge. That is, candidates who are able to provide a completely accurate explanation of their response receive the highest score point. Those who provide an accurate response with an incomplete or inaccurate explanation receive only partial credit. In our early work, we had begun to design separate scales to differentiate candidates' conceptual and procedural knowledge. Ideally, we might have conducted factor analyses to test construct validity and design appropriate scales for each construct. We were reminded, however, of time constraints and the intended use of the assessment and we followed the advice of the CSDE Title II Project Director in constructing a scoring protocol that would ensure reliability and facilitate scoring as a component of the unit's assessment system for NCATE accreditation.

Previous research has demonstrated the challenges associated with separating the constructs of conceptual knowledge and content knowledge. Researchers at Stanford University conducted a study of the validity and usefulness of mathematics achievement tests by distinguishing items that assessed mathematical content knowledge from those assessing mathematical reasoning (Kupermintz et al., 1995). In a companion study, researchers studied science achievement tests and distinguished those items that required mostly scientific reasoning around elementary or every day science content from those items that required more formal and advanced science content knowledge (Hamilton et al., 1995). In mathematics, items that assessed reasoning included ranking decimals and fractions by size order, solving word problems requiring logical inference, and evaluating statement inferred from a word problem with fractions. In science, items assessing reasoning required nonformal content knowledge or knowledge of everyday science. Such items included reading and interpreting graphs, making inferences from given scientific facts, and identifying the basis of statements made about scientific phenomena.

The Stanford University achievement test studies were valuable resources for us as we developed and revised our items. For example, we intentionally designed items that required minimal content knowledge and emphasized reasoning. We also focused on content commonly taught at the elementary level. For example, understanding the concept of rational numbers forms the basis for most elementary mathematics (Behr, Lesh, Post, & Silver, 1983). This concept includes part-whole relationships, quotient, number lines, and decimals. To assess understanding of the concept of rational numbers, our assessment included items testing fractions, mixed numbers, decimals, number sense, and operations.

In science, the selection of content was more challenging because of the different content strands. We began by identifying six strands (Harlen, 1993). See Appendix A3 for documentation of our early work. Ultimately, we reduced the strands to three (life, physical, and earth/space sciences) based on the *INTASC and National Science Education Standards*.

Thus, our purpose in selecting content was not to be inclusive but exclusive. That is, because our purpose was not to assess the complete range of candidates' content knowledge, but rather to assess candidates' **conceptual knowledge**, we used the literature and national and state standards to ensure that assessment of conceptual knowledge was embedded in content commonly taught at the elementary level. We intentionally excluded items that required knowledge of content beyond what would be most commonly taught at the elementary level. In particular, we reviewed the Connecticut Curriculum Frameworks and aligned items to the Frameworks (see for example, Section 2.VI).

During the course of our project, there was some discussion about designing items around common misconceptions. Our review of the literature on misconceptions revealed that the literature is much more consistent in identifying misconceptions in mathematics than in science. This may be because of the variety of scientific disciplines (life sciences, physical sciences, earth/space science). More research needs to be conducted in each of the scientific disciplines to replicate findings across studies. We also found a wide range of lay literature around science misconceptions, in addition to studies published in peer reviewed journals. The review of the literature encouraged us to reaffirm our original intention of assessing

candidates' conceptual understanding in a more general sense, rather than knowledge about specific content that some research has found likely to be misunderstood by children and adults. By embedding our items in basic content that is more likely to be understood (rather than misunderstood), we provided candidates more opportunity to demonstrate their conceptual understanding of basic scientific concepts they were likely to "know" as facts yet not necessarily "understand."

The assessments were designed to measure candidates' conceptual knowledge at the start of their program. The assessments are intended to provide candidates information about their strengths and weaknesses as well as provide us information to enhance program design. That is, we plan to develop elementary education content courses, online modules, and/or components of our mathematics and science strategies courses that will address candidates' weaknesses. We believe that our science and mathematics strategy courses currently build teachers' conceptual knowledge. For example, science and mathematics strategy courses are taught in ways that promote scientific inquiry and mathematical reasoning. Ultimately, we plan to refine the assessments and pre and post test candidates before and after they take the strategy courses to determine the degree to which candidates strengthen conceptual knowledge after participating in the learning experiences afforded by the strategy courses.

Because this grant project focused on the development and validation items, rather than an evaluation of our candidates' performance, we informed candidates that the results of the assessment would not be recorded as part of their program (See Appendix B for a description of the information shared with students prior to administration of the pilot. We did, however, share results with the candidates as information about their strengths and weaknesses compared to the group average and in each subcategory (see Appendix B for a Sample Student Report).

Reliability

After the rubrics were developed, the principal and co-principal investigator scored the assessments and illustrated the rubrics with actual student responses as benchmarks for each score point. Assessments from three students in each content area were scored again and reviewed for consistency. These assessments then became “anchor” papers for the scoring sessions. During the first scoring session, the rubrics were reviewed and scorers were asked to score the responses from the three anchor papers. Scores were discussed and explained. Then scorers scored the remaining assessments, completing a “score card” for each assessment they scored. Each assessment was scored by at least two raters, typically one of the scorers and one of the principal investigators. Whenever there was a discrepancy, scores were discussed until agreement was reached. A matrix was developed to assist us in reviewing items that were most often scored inconsistently. These items were reviewed for clarity and were revised or eliminated in the final item pool (see Appendix B for an example of a score card and inter rater reliability matrix).

Scoring Protocols

Initially, in science, three scales were constructed (factual knowledge, conceptual knowledge, and scientific inquiry knowledge). Our intent was to separate the constructs assessed in each item. In mathematics, we developed two scales, one with two score points for dichotomous questions, and one with three score points for questions that required elaborated explanations. We consulted with one of our team members, who has validated and published assessments. With her assistance, we revised the mathematics rubrics and developed one three point scale applicable to all questions.

After discussion with consultants, we revised and simplified our science rubrics. The final science scoring rubric consists of one three point scale similar to the mathematics scale.

We developed the rubrics after we piloted the assessments. Therefore, we used actual student responses to inform our rubric development, which was a recursive process. That is, we drafted a rubric, illustrated the score points with responses, revised our criteria and descriptors as the sample responses provided us more information about descriptors, and so forth. We noted items where responses consistently indicated that the items might have been unclear. Later, we revised these items or eliminated them from the final item pool.