Santa Clara County Office of Education

Protocol for Analyzing Math Benchmark Assessments
April 25th, 2012

Math Team: Exemplary Team **Grade Level:** 7th **Math Subject:** 7th Grade Math

Introduction:
Our team seeks to improve its upcoming Math Benchmark Assessments in June. In order to accomplish this work, grade level teams of Math teachers will participate in an instructional session where they will learn how to use item analysis, classical statistics, and Wright Maps to analyze the June, 2011 Math Benchmark Assessments. Teacher teams will use the tools and reports found on the SChoolPlan system to conduct the analyses, generate finding, and make recommendations for the improvement of the June, 2012 Math Benchmark Assessments.

Purpose:
To improve the upcoming June, 2012 Math Benchmark Assessments through a careful analysis of the previous June, 2011 Math Benchmark Assessments using item analyses, classical statistics and Wright Maps.

Deliverables:
- Key findings about the June, 2011 Math Benchmark Assessments based on item analysis, classical statistics, and Wright Maps.
- 5-10 Recommendations for improving the June 2012 Math Benchmark Assessments
- 3-4 Action Steps the team will take to implement the recommendations.

Description:
Each team of teachers will receive copies of the appropriate June, 2011 Math Benchmark Assessment. Teams will review the actual assessments and answer the following questions.

Theory of Action:
Step 1: After reviewing the assessment items, does there appear to be a theory of action that underlies the overall assessment?
- One theory of action might be that the assessment uses items aligned to standards and built on Bloom's taxonomy starting with items based on knowledge, comprehension, and moving to items based on analysis.
- A second theory of action might include items that are based on student recognition of key math ideas moving to a conceptual understanding and ultimately leading to application of the concepts to the solution of problems.
- A third theory of action might include items that involve single step, two step and multiple step solutions.

- A fourth theory of action, might state that some items are aligned to more challenging standards than others.

Findings:

- *There are several potential theories of action working within this assessment. The assessment is aligned to three math areas including Algebra and Functions, Measurement and Geometry, and Number Sense. A prerequisite for students to do well in Algebra is built upon their Number Sense. Thus it would be expected that assessment items aligned to Number Sense would be less challenging to students than items aligned to Algebra 1.*
- *The assessment includes items that require students to use to solve procedural mathematical problems (problems 1-8 are examples) and there are also problems that require students to apply their knowledge of math skills to solve problems. (Problems # 11 and #12). It would be expected that application problems would be more challenging for students than pure math problems.*

Step 2: If there is no explicit theory of action that underlies the assessment, review the items and identify those items that may be more challenging for students to solve based on Bloom's taxonomy or other criteria for complexity that the team identifies.

Findings:

- *Some of the problems on the test require students to use multiple steps to solve them. (Problems 11, 12, 22).*
- *The test includes items aligned to Number Senses and Algebra 1. It would be expected that Algebra 1 items appear to be more challenging than items aligned to the topic of Number Sense.*

Step 3: Review the items from the perspective of item quality. Review the document that describes the quality characteristics for selected response items. Do any of the items reflect quality that do not meet the criteria and may be difficult for students?

Findings:

- *Make sure the question you're writing matches the standard or skill description.*
    - *Item 20 is aligned to Number Sense 2.0 in that it uses exponents in working with fractions, it also is aligned with Algebra and Functions 1.1 in that it uses variable to represent an expression. It would be difficult to gauge from this item that it provides information about students' knowledge of Number Sense 2.0 or Algebra and Functions 1.1.*

- *Ensure consistency of style, format, text, and graphics within items and subject areas.*
    - *There is no consistency in the formatting of the stem for question 21 when compared with other items on the assessment.*

Summary:
Based on the development of findings from a review of the actual assessment, identify items on the test that may be more challenging for students to solve based on the criteria that your team developed or based on quality issues with the items.

Challenging Items Based on Key Criteria	Challenging Items Based on Quality
Items 11 and 20 may be more challenging because they require the application of math ideas to real world problems. *Items aligned with number sense (17, 19, 20, 21, 23, 24) should be less challenging for students than items aligned with Algebra and Function. 1, 2, 3, 4, 5, 9, 10,11, 12, 13, 14, 16, 16, 22)*	*Item 20 appears to have alignment to both Number Sense and to Algebra and Functions.* *Item 21 lacks consistency in formatting the stem.*

Based on this initial review of the items, identify several recommendations for improving the assessment.

- *Ensure that there is a balance of items with the application of math procedures balanced with items that require procedures only.*
- *Improve the alignment of item with Number Sense or Algebra and Functions. Include additional items that ensure students can achieve Number Sense 2.0 with rational numbers before using an item that requires both Number Sense knowledge and knowledge of Algebra and Functions.*
- *Improve stem formatting for item 21.*

Item Analysis:
Your packet will include an Answer Frequency Report that is an analysis of items that includes the percentage of students who selected the correct response and the percentage of students who selected the distractors for each item on the test. The items will be organized from the items with the lowest percentage of correct responses to the items with the highest percentage of correct responses.

Step 4: Review the items with the least percent correct. Do these items reflect possible quality issues? Do the items reflect the challenging criteria previously described in steps 1 and 2?

Findings:

- *Only 41% of the 386 7^{th} grade students who picked item 17 selected the correct answer, B, that it is a composite number. 51% selected the incorrect response A that it is a prime number.*

- *Items 11 and 12 were hypothesized to be challenging for students because they involved the application of ideas and they also required multiple steps to solve. The item analysis appears to not validate this hypothesis as 93% of students got problem 11 correct and 76% of students got problem 12 correct.*

Step 5: Review and circle distractors with greater than 40% response rates. Do these distractors represent potential student misconceptions or errors in solving the problem?

Findings:

- *Only 54% of students got item 20 correct (Distractor D) with 24% of students selecting the incorrect distractor (Distractor B). Quality issues with this item have been previously described. Selection of distractor B may reflect an error in the ability to evaluate fractions where the variables used have larger exponents within the denominator.*
- *Only 60% of students were able to select the correct response for item 14. This could reflect a quality issue in the size and clarity of the graphic used for this problem.*

Summary:
Based on the development of findings from the item analysis, identify items that may be challenging for students or represent quality challenges.

Challenging Items Based on the Item Analyses	Challenging Items Based on Quality
Questions 11 and 12 were hypothesized to be challenging but proved not to be challenging based on the correct response percentages for students.	*Based on the item analysis, items 14 and 20 may have issues with quality.*

Based on this item analysis review of the items, identify several recommendations for improving the assessment.

- *Improve the rigor for application items like 11 and 12.*
- *Improve the quality of the graphic for item 14*
- *Ensure that that there are sufficient items aligned to individual standards before providing students with more challenging problems that involve one or more standards for their solution.*

Classical Statistics Analysis:
The Answer Frequency Report also provides classical statistics data on the items. Please use these classical statistics analysis to generate findings and recommendations below.

Step 6: Review the classical statistics document from the perspective of the Point Biserial statistic. This statistic differentiates items based on a consistent level of difficulty. Students who perform well on the overall test should get easier items correct a high percent of the time and more challenging items correct 50% of the time. Students who do not do well overall on the test may get the easier items correct 50% of the time, but will get the more challenging items correct at a much lower percent. A positive point biserial number indicates that this rule is generally followed throughout the test for the test takers. A negative point biserial means that there is inconsistency in this rule and that the item should be reviewed for quality. A positive point biserial of at least .015 is recommended. Good items are considered to be above 0.25.

Identify items that have a negative point biserial and review them for quality and then record your findings about these items.

Findings:

- *None of the items demonstrated a negative point biserial value. However, item 17 demonstrated a 0.13 point biserial value for all 7th graders who took the test.*

Step 7: Review the Answer Frequency Report from the perspective of Item Difficulty. The Scale Score Difficulty statistic reflects the probability of getting an item wrong converted to a CST scale score metric. Items with high scale scores are very difficult; there is a high probability of getting them wrong. Items with low scale scores are easy, with a low probability of getting them wrong. Items that have very high difficulty scores or very low difficulty scores should be checked to determine if it is the quality of the item that is making them difficult or the theory of action that drives the assessment

Identify items that have very high difficulty scale rankings or very low difficulty scale score rankings and record your findings about these items.

Findings:

- *Item 17 demonstrated a level of item difficulty of 410 which is a very high scale score. The next highest scale score level of difficulty was item 10 with a scale score difficulty level of 337.*

Step 8: Review the Answer Frequency Report from the perspective of Item Reliability. The Item Reliability statistic ranges from 0.0 to 1.0, where 1.0 means "perfectly reliable". It is a measure of how well the item is able to reveal differences between high performing students and low performing students. Two quantities go into item reliability -- a) the spread of the students on the item (as measured by their differing probabilities of success on the item); and b) the average margin of error around each student's probability of success. When the students are well-spread out and the average margin of error is very low, the item reliability approaches 1.0. When the students are clumped together in the middle of the scale, or when the average margin of error is very high, the item reliability approaches 0.0. When items are given an item reliability of 0.0 or near 0.0, it means that the average margin of error is the same as or greater than the spread of the students along the scale. In other words, viewed through the lens of that item, the students are one big blur and it would be difficult to attribute their performance on the item to the learning target to which it was aligned.

Identify items that demonstrate reliabilities at 0.0 or near 0.0 and record your findings about these items.

Findings:

- *Item 17 demonstrated an item reliability of 0.0.*
- *The next lowest level of reliability was item 20 with a reliability value of 0.33.*

Summary:
Based on the development of findings from the classical statistics analysis, identify items that may be challenging for students.

Challenging Items Based on the Classical Statistics Analyses	Challenging Items Based on Quality
From the perspective of the point biserial, item reliability, and item level difficulty, item 17 presented challenges based on the statistics for this item that could be related to the ability of students to identify prime or composite numbers from larger double digit numbers like 91.	*From the perspective of the point biserial, item reliability, and item level difficulty, item 17 presented challenges based on the statistics for this item that are probably due to item quality issues.*

Based on this Classical Statistics review of the items, identify several recommendations for improving the assessment.

- *Investigate both the quality of item 17 in terms of the level of difficulty of numbers that it presents to students. It may be appropriate to include smaller single and double digit numbers for students to evaluate as prime or composite.*

Wright Map Analysis
Your packet contains a Wright Map of the assessment items. A Wright Map places the items on the same CST scale as the students. There is an expectation that the spread of students should more or less match the spread of students on the Map. There should be items within performance bands that also align with students at that band. Students will have a 50% probability of getting items correct that are in their band. Please use the Power Point Presentation that describes how to use the Wright Map to analyze assessment items. You can also find detailed descriptions of how to use Wright Maps to analyze assessment items in the book called "Three Facets of Formative Assessments".

Step 9: Based on your review of the Wright Maps, does the distribution of students match the distribution of items based on the level of item difficulty? Please describe.

Step 10: Based on your review of the Wright Maps, does the Map corroborate or not corroborate hypotheses that you previously made about item complexity or quality? Please describe.

Finding:

- *Most of the items on the Wright Map do not match the distribution of student performance on the assessment indicating that many of the items are not challenging enough for students.*
- *Some of the number sense items such as 17 and 20 score at higher levels of difficulty than would be predicted based upon the theory of action. Item 17 which is categorized as a Number Sense Item scores in the proficient range higher than any other items on the test.*

Step 11: Does the Map identify new items that need to be reviewed because they are easier or more challenging than expected?

Findings:

- *Items 11 and 12 that were hypothesized to be challenging to students actually scored at a low proficiency level on the Wright Map. Item 11 scored at the Far Below Basic Level and Item 12 scored at the Below Basic Level.*

Summary:
Based on the development of findings from the Wright Map analysis, identify items that may be challenging for students.

Challenging Items Based on the Wright Map Analyses	Challenging Items Based on Quality
Items 11 and 12 performed at lower levels of item proficiency than would be expected for items that involve the application of math skills.	*There may be a quality issue with item 17 in selecting only larger double digit numbers to evaluate student ability to differentiate prime and composite numbers.*

Based on this Wright Map review of the items, identify several recommendations for improving the assessment.

- *Improve the rigor of items that require application of math content and skills. Possibly consider that they all involve multi-step solutions.*
- *Review the quality of item 17 and ensure that there are more less challenging opportunities for students to identify prime and composite numbers.*

Based on this review of the 2011 7th Grade 2nd Quarter Math Benchmark Assessment, the team developed the following Action Steps to improve the quality of the assessment.

Action Step 1: The team will collaborate in revising the assessment to include more items that involve application and also require at least 2 steps to solve the problems.

Action Step 2: The team will collaborate in revising items that demonstrated poor quality as identified above.

Action Step 3: The team will collaborate to build several problems that elicit student understanding of prime and composite numbers.

Action Step 4: The team will collaborate to build attractive distractors into at least 3 items that will elicit error patterns in student thinking.

References

Dan Mason, Mark Moulton, Ph.D., Dale Russell, Ed.D., Diana Wilmot, Ph.D., 2009. Three Facets of Formative Assessment. Santa Clara County Office of Education. San Jose, CA.