

Changes in Assessments and Assessment Systems Since 2002



SCOTT F. MARION
NATIONAL CENTER FOR THE IMPROVEMENT OF
EDUCATIONAL ASSESSMENT
NRC BEST PRACTICES IN STATE ASSESSMENT:
WORKSHOP I
DECEMBER 10-11, 2009

Many changes in tests and assessment systems:

2

- Grades tested
- Form design
- Item type
- High school assessment
- Use of interim assessments
- Reporting systems
- Alternate and ELL assessment
- State DOE capacity and resources



Grades tested

3

- **IASA (1994-2001) required testing at one grade each elementary, middle, and high school**
 - All states eventually fulfilled this requirement (but not by 2001)
 - At least 12 states exceeded this requirement
 - ✦ Many states that exceeded IASA did not all test every subject every year
- **NCLB required testing 3-8 plus high school by 2005**
 - All states have met this requirement
 - Many states also use multiple high school tests (EOC)



Implications of increase in grades tested

4

- **Content standards**
 - Necessary to move from grade-span to grade-level expectations
 - ✦ Coherence (or lack thereof) of content expectations across grades
- **Achievement (performance) standards**
 - Moving from “just so” stories to cross-grade articulation
- **New accountability models**
 - Proliferation of growth and VAM models



Form Design

5

- **2000-2001**
 - At least one state (MSPAP) used full matrix design
 - Several states used matrix/common design with matrix counting in at least school score and in some cases in the student score
- **2008-2009**
 - Few if any states use full matrix or even count matrix items in “school results”
 - Efforts being made to shorten operational forms
- **Implications**
 - Narrowing (or focusing, depending on your perspective) of curricular goals
 - ✦ Even thinner sampling of remaining domain
 - A shift in focus of form design from school to student
 - Harder to include “memorable” task because of equating challenges (less of a problem with matrix designs)



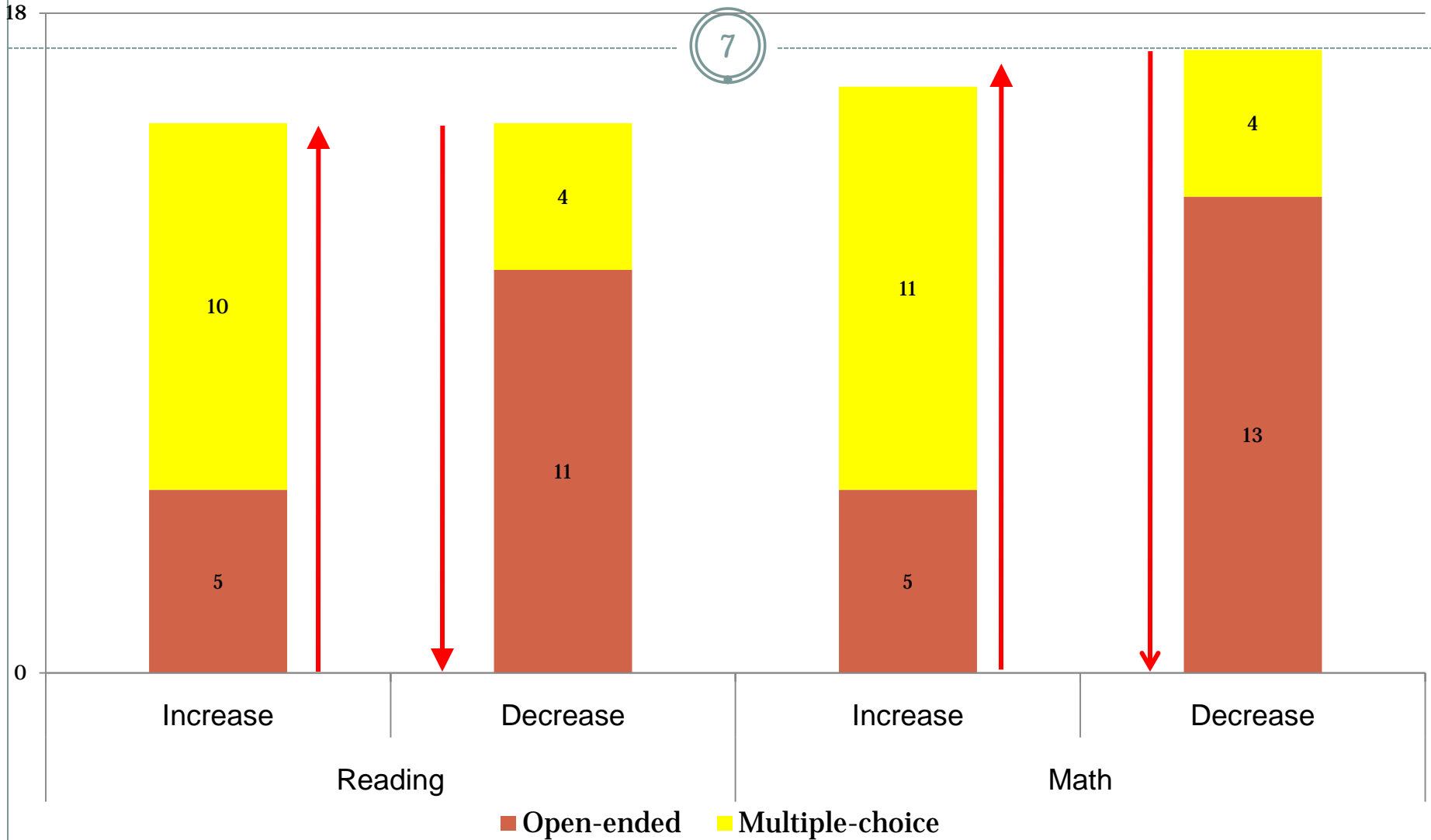
Item Types

6

- *There is no need to include item types other than multiple choice on NCLB assessments*
 - Susan Neuman, former Assistant Secretary of OESE to a [stunned] lunch crowd of approximately 1000 testing people at CCSSO's large-scale assessment conference in 2002.
- There is little question that there has been a shift away from complex performance assessments toward multiple-choice items during the past nine years
- Many states still use some form of “open-ended” items, but this term hides the important differences between 2-point short constructed response items and extended, complex performance tasks



Number of states reporting increase/decrease in open-ended/multiple choice items types 2002-2009 (GAO, 2009, 47 states responding)



“Open-ended” *Tasks*: Before & After

8

- First slide is from 1996 CAPT (CT)—integrated lengthy task involving group work, multiple activities (4 pages of tasks), extensive materials (16 pages of source materials)
 - Could have also included examples from MSPAP, KY, WY, CA and others
- Second slide includes two 2009 MCAS constructed-response items from grade 8 mathematics
 - Good items, but clearly more limited—these are characteristic of today’s “open-response” items
- Third slide also from 2009 MCAS showing that some still include extended response tasks in content areas other than writing (with a few pleasant exceptions in science)
 - As good as these tasks are, still much more constrained than the types of tasks we were seeing in the 1990s



1996 CAPT Interdisciplinary Assessment

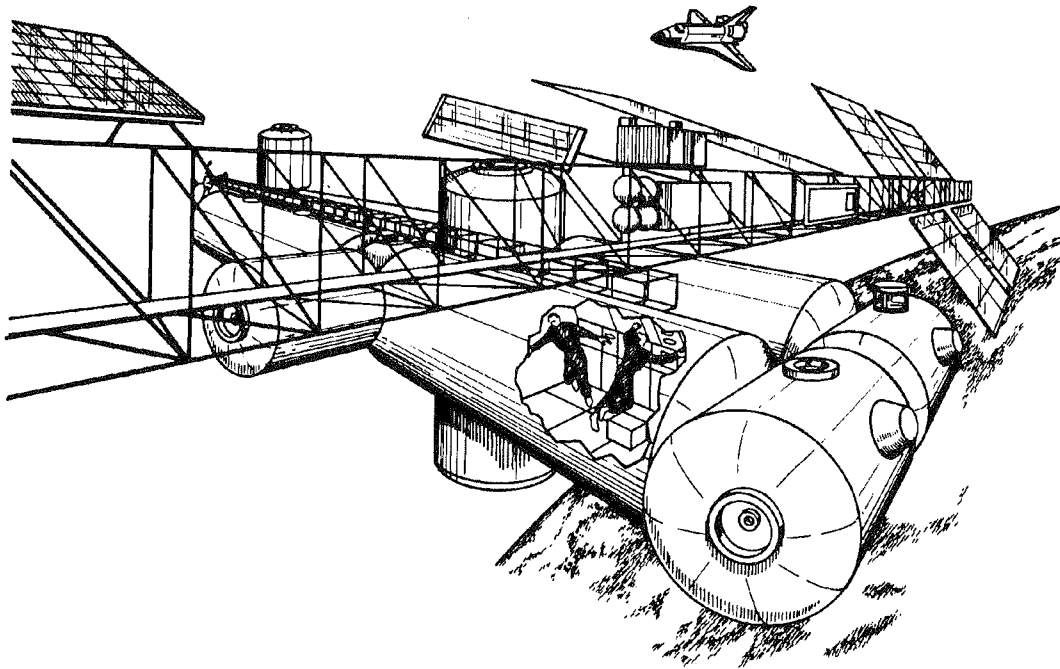
Space Station

About this Assessment

In this interdisciplinary assessment you will think about and respond to an important issue—the development of a space station. You will discuss the issue with your classmates, read and evaluate several sources of information related to the issue, and write a speech in which you take a position on the issue. In working on this assessment, you will use skills and knowledge you have learned in your language arts, mathematics, science, social studies and other classes.

The Issue

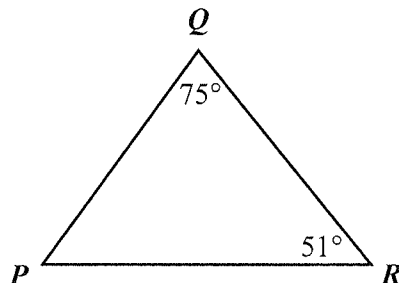
In recent years considerable controversy has arisen over the U.S. funding of the development of a space station. Because of the huge federal debt and other national needs, some people believe that money could be spent in better ways. Others believe that the development of the space station is vital in maintaining the United States' position as a leader in technology.



You will begin the activity with a brief, 10-minute group discussion. The group discussion will help you start thinking about this issue. Your teacher will arrange the class into small groups of three or four students each.

Questions 9 and 10 are short-answer questions. Write your answers to these questions in the boxes provided in your Student Answer Booklet. Do not write your answers in this test booklet. You may do your figuring in the test booklet.

- 9 The figure below shows $\triangle PQR$ and two of its angle measures.



What is the degree measure of $\angle P$?

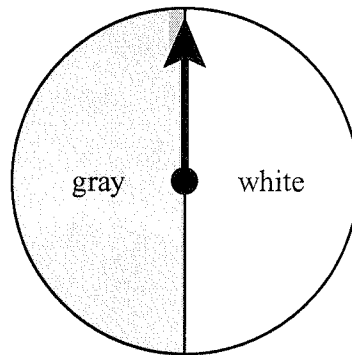
- 10 James drove 200 miles in 4 hours. At this rate, what is the total number of hours it will take him to drive 150 miles?

Question 11 is an open-response question.

- **BE SURE TO ANSWER AND LABEL ALL PARTS OF THE QUESTION.**
- **Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.**
- **If you do the work in your head, explain in writing how you did the work.**

Write your answer to question 11 in the space provided in your Student Answer Booklet.

- 11** Daniel has a spinner divided into two congruent sections, as shown below.



- Daniel will spin the arrow on the spinner one time. What is the probability that the arrow will stop in the gray section? Show or explain how you got your answer.
- Daniel will spin the arrow two times. What is the probability that the arrow will stop in the gray section both times? Show or explain how you got your answer.
- Daniel will spin the arrow three times. In your Student Answer Booklet, construct a tree diagram that shows all the possible outcomes that can occur.
- Based on your diagram from part (c), what is the probability that the arrow will stop in the white section **at least** one time when Daniel spins the arrow three times? Show or explain how you got your answer.

Item Type **Implications**

10

- **Increased focus on breadth and decreased focus on depth**
 - Expectations for student learning—as conveyed by the types of items/tasks used—are notably less complex
- **Signaling that teaching involves making sure that students have had an opportunity to learn a broad array of content and skills even if coverage is superficial**
- **Sends the message that efficacy and “objectivity” are more important than complexity, integration, and multiple appropriate answers**
- **Do “21st century skills” or college/work ready knowledge and skills focus on the types of processes students use to solve multiple-choice items, especially to the exclusion of the processes required for complex performance assessments?**



High School Assessment

11

- In 2008 13 states, compared with 10 in 2003, were using end-of-course exams as part of their state assessment systems (CCSSO, 2008)
 - Now that USED has indicated a bit more willingness to consider EOC exams, more states may shift in this direction
- Use of college entrance tests as at least part of high school assessment system
 - Required census testing of the ACT in 5 states (IL, CO, KY, MI, WY)
 - Maine's high school assessment is the SAT
- Graduation exams appear to be on a slight increase since 2002, but many are seeing the bark to be worse than the bite



High School Assessment **Implications**

12

- **EOC vs. survey**
 - Curriculum sensitivity compared with potential for measuring transfer
 - ✦ This could be a legitimate value decision if survey tests were actually designed to measure transfer of important concepts and big ideas, but unfortunately this is rarely the case
 - ✦ Given current testing time constraints, EOC tests should allow testing concepts in greater breadth and depth than with survey tests
- **Use of college-entrance exams (which are now mysteriously called “college readiness” exams)**
 - Could help reveal potential of certain students
 - Could narrow the curriculum even further compared to survey tests designed to measure the full range of standards



Interim Assessment

13

- In spite of concerns about over testing with the shift to 3-8, HS testing requirements, districts and schools have chosen to add another form of testing to the mix
- Interim assessments have proliferated in recent years as documented by the following information from two of the leading interim assessment providers:
 - NWEA has seen a growth in the number of districts using its assessment system from just under 400 at the end of 2001 to 4,200 now, an increase of 950% (*R. Yeagley, via email 11/30/09*)
 - Renaissance Learning has seen about a 45% growth in the number of schools using its STAR assessment systems [just one of its products] since 2001 (*J. McBride, via email 12/4/09*)



Interim Assessment **Implications**

14

- We are data rich....but information poor
- Limited research regarding the assessment specifications, human capacity, and support structures necessary for interim assessments to lead to improvements in student learning
- Tremendous variability and essentially no oversight in the quality of the products being offered
 - While peer review is not perfect, it does serve as at least one check on the quality of state assessments (also have TACs, etc)
- An over-reliance on poor quality multiple-choice items, but little capacity of customers to tell the difference
- Yet, district leaders tend to like interim assessments...
 - Fast results
 - Positive stories
 - Belief in positive uses
 - Local control



Reporting Systems..some good news

15

- **There has been noticeably more attention to reports and reporting systems since 2002**
 - Perhaps due to increased reporting/notification requirements
 - Perhaps due to better data systems
 - Certainly due, in part, to better utilization of newer technology
 - ✦ See www.SchoolView.org for a great example of where reports are heading
 - ✦ Careful and intentional designs to tell the most important stories and to present data in ways that easily can be turned into information, decisions, and hopefully appropriate actions



Assessing Special Populations... A bit more good news

16

- **AA-AAS**
 - Major shift in content from functional to academic
 - Learning to find the appropriate balance between standardization and technical quality
 - Making great strides in evaluating and documenting technical quality
 - ✦ Some of our best examples of validity arguments in practice
- **AA-MAS**
 - Pushing our understanding of construct comparability
 - Improving our techniques for minimizing construct-irrelevance
 - Reinforcing the notion that assessment cannot fix an instructional problem
- **ELP**
 - Challenge of assessing multiple domains
 - Improving understanding of patterns of language acquisition
 - Improving understanding of the relationship between language and academic proficiency
 - Beginning to learn about evaluating technical quality



State DOE Capacity & Resources..not such good news

17

- GAO reports, Tom Toch's *Margin of Error*, a number EdWeek articles have documented the loss of relative capacity in state assessment leadership
 - Most states have seen at least a three-fold increase in the number of tests (don't forget alternate and ELP!), yet have seen nowhere near that increase in personnel
 - Further, except in the larger states, assessment personnel are also now the accountability personnel
 - Almost every state is suffering significant fiscal woes
- **Implications**
 - Shift in expertise from DOE to vendors
 - Less capacity to monitor and "partner" with vendors
 - Less time to think about innovation
 - QA/QC of performance assessment requires work beyond what is necessary for multiple-choice items (e.g., benchmarking, monitoring scoring)



Some reasons for the changes

18

- **Purposes and Uses (accountability)**
 - NCLB required a rapid turnaround of results
 - AYP based on a “head-counting” methodology
 - Shift to ACT/SAT designed to increase relevance and promote college access
 - Interim assessment—searching for any way to improve (and co-opting formative assessment research)
- **Technical quality oversight**
 - Peer review guidance
 - ✦ Alignment has been privileged and criteria have become reified
 - ✦ Lack of understanding of comparability and within-year equating
 - ✦ Validity has received far too little attention
- **Cost and capacity**
 - NCLB supported some development costs, but not nearly enough to offset increased ongoing operational costs
- **Perceived value**
 - Some have argued that since multiple-choice and open-ended scores are “well correlated,” open-response tasks are not worth the cost, but...
 - Correlation is not validity and many argue that the unique contributions of open-response tasks are more important than the overlap



For more information...

19

- smarion@nciea.org
- www.nciea.org

