A Study of "Irrelevant" Items: Impact on Bookmark Placement and Implications for College and Career Readiness[1]

Susan Cooper Loomis
Assistant Director for Psychometrics
National Assessment Governing Board[2]

Abstract
The judgmental standard setting studies for 12th Grade NAEP Preparedness Research are the focus of this paper. While the standard setting methodology was very familiar, the post-secondary context of standard setting was new to the Governing Board. This was the first effort to set a cut score on NAEP to represent minimal academic preparedness, and the studies were to be for both college course placement and entry to job training programs. The Governing Board made no assumptions regarding the relationship between preparedness for college course work and the job training programs; the Board wanted to know if the academic knowledge and skills required to be minimally prepared for "college and career" are similar. Clearly, there was much to be learned about standard setting in this new post-secondary context, and this paper reports on the issues surrounding "irrelevant items" in a standard setting context.

The potential for a mismatch of academic knowledge and expertise on the part of occupational job trainers with that measured by the grade 12 National Assessment of Educational Progress (NAEP) was anticipated, especially with respect to the grade 12 mathematics NAEP. The compilation of results from the studies implemented across a two-month period pointed to an ultimate conclusion that there should be no reference points on NAEP associated with the outcomes of these studies. The results of the judgmental standard setting studies simply did not confirm nor were they confirmed by results from other studies. Gathering more information about the judgments and potential impacts on the judgments that were the ultimate outcomes of the process was a growing research priority. A special study was designed and implemented to collect more information about the issue of irrelevant items and the potential impact of these items on the judgments of panelists.

Introduction and Background for this Research

The need for U.S. students to be academically prepared for college and careers seems clear to everyone from the President to the postman. College and career readiness research is at the forefront of major large-scale assessment programs at the national and state level in this country. The Common Core State Standards call for measures and standards of college and career readiness; ACT, Inc. has college and workplace readiness benchmarks for their assessments to predict college course grades; the College Board has developed benchmarks to predict freshman year grade point averages; and the National Assessment Governing Board has launched an

[1] I wish to acknowledge the assistance of Dr. Luz Bay and Mr. Chris Clough of Measured Progress in preparation of this paper. They provided data collected in the special study for use in writing this paper.
[2] This paper was prepared for presentation in the session *Setting Academic Preparedness Standards for Job Training Programs: Are We Prepared?* at the annual meetings of the National Council on Measurement in Education, Vancouver, British Columbia, Canada, April 14, 2012. More information about the 12th Grade NAEP Preparedness Research is available at www.nagb.org.

extensive program of research to identify preparedness levels for the grade 12 National Assessment of Educational Progress in reading and mathematics. Additionally, some states are establishing their own post-secondary readiness/preparedness indicators.

The work of the National Assessment Governing Board to report preparedness of 12th graders on the National Assessment of Educational Progress (NAEP) has been underway for almost a full decade. The Governing Board was urged to explore the feasibility of using NAEP for reporting "12th grade students' readiness for college-credit coursework, training for employment, and entrance into the military" (National Assessment Governing Board, 2004, p. 6) as a way to increase interest and participation in the 12th grade NAEP. This recommendation was studied by the Governing Board for the first few years through a series of commissioned papers, the work of Ad Hoc Committees, and finally appointment of a technical panel in 2007 to advise the Governing Board in developing a program of 12th grade NAEP preparedness research.

Importantly, the focus of the Governing Board's program of research is on *preparedness,* as opposed to *readiness.* This terminology was chosen to clarify that the research is related only to the academic knowledge, skills, and abilities that can be assessed by NAEP. "Readiness," as it is generally used, connotes the addition of behaviors and personal characteristics that are so important to success in all aspects of post-secondary life. But, those are not and cannot be part of the assessments in NAEP. Further the Governing Board resolved to have research results guide their conclusions regarding the relationship between preparedness for college and job training in occupations; they made no assumptions that the same level of academic preparedness is required for both post-secondary choices. These are two distinguishing characteristics of the Governing Board's approach to the analysis of student preparation for post-secondary activities.

The Technical Panel for 12th Grade NAEP Preparedness Research[3] worked for about 18 months to produce a comprehensive research plan for the Governing Board (National Assessment Governing Board, 2009). The general approach recommended by the Technical Panel was to use a variety of methodologies across several studies to determine if mutually confirmatory evidence exists to support the validity of using 12th grade NAEP for reporting preparedness. The studies were selected to provide information that allows each potential reference point or range of score points on the NAEP scale to be evaluated relative to points or ranges derived from other studies. Their recommendations were adopted by the Governing Board in 2009, and research studies were designed to carry out this general plan. Research studies implemented for the 12th Grade NAEP Preparedness Research include:

- content alignment of NAEP in reading and mathematics with ACT, SAT, ACCUPLACER, and WorkKeys;
- statistical linking of NAEP reading and mathematics with SAT critical reading and mathematics, and statistical linking of NAEP reading and mathematics with a state representative sample of Florida 12th graders who took NAEP in 2009 and had taken the SAT, ACT, or ACCUPLACER, and analysis of the statistical relationship of performance on NAEP with college freshman-year performance of the Florida students who went to a

[3] Members of the Technical Panel included experts in high school-college transition, psychometrics, NAEP technology, industrial-organizational psychology, validity, two-year colleges, military and civilian job training classification systems, and research design.

public post-secondary institutions in the state following graduation from high school in 2009;
- a benchmarking study of Texas college students;
- a nationally representative sample of higher education institutions regarding tests and scores used for placing first-year college students in credit-bearing courses and in remedial courses in reading and mathematics; and
- judgmental standard setting studies to set a score on NAEP to represent minimal academic preparedness for college course work or entry in job training programs.

Meanwhile, during the early stages of the Governing Board's deliberations regarding 12th grade preparedness research, a new framework for NAEP reading in grades 4, 8, and 12 had already been developed and scheduled for implementation, and a new framework for grade 12 NAEP mathematics had been implemented in 2005—complete with new achievement levels. The Governing Board partnered with Achieve to evaluate the frameworks with respect to the grade 12 requirements and to recommend modifications that might be needed to have the 12th grade NAEP in both reading and mathematics serve as measures of preparedness for post-secondary activities in both higher education and the workplace—both civilian and military. Achieve appointed panels of content experts to conduct this review and make recommendations for modifications (Achieve, 2006a and 2006b). The recommended modifications to the frameworks were made, and items were developed according to that framework for the 2009 NAEP in reading and mathematics.

Judgmental Standard Setting Studies

Overview

No national or systematic set of standards could be identified for workplace preparedness, nor was it possible for the Governing Board to develop a partnership for linking NAEP with any civilian or military assessments of job-place "readiness." Thus, the Technical Panel concluded that judgmental standard setting was the only potential study design for developing indicators of minimal academic preparedness for the workplace job training programs. A combination of lack of data in this area and an abundance of technical expertise on the Technical Panel led to rather detailed recommendations regarding the criteria for selection of job training occupations to study, types of panelists (job trainers) to recruit for the workplace preparedness studies, and study design elements to include replicate panels. An important recommendation regarding the selection of occupations to include in the studies is that at least three months of post-secondary job training and less than a bachelor's degree is required. The Governing Board's definitions of preparedness used in this research are as follows:

- College preparedness is the minimal academic knowledge and skills needed for placement in a college level course of the sort that fulfills general education requirements without remediation.
- Preparedness for career job training is the minimal academic knowledge and skills needed for entry into a job training program in the occupation.

If a college degree is required for entry into the occupational area, there should be no difference in the minimal academic requirements.

A modified bookmark method was selected for the judgmental standard setting work. ACT, Inc. had implemented Mapmark, which is a modified bookmark procedure, for setting NAEP achievement levels for grade 12 mathematics in 2005 (ACT, Inc,. 2005), for grade 12 economics in 2006 (ACT, Inc., 2006), and for grades 4, 8 and 12 science in 2009 (ACT, Inc., 2009). Based on those previous study designs, the Governing Board developed a design document to be used for the judgmental standard setting preparedness research studies. The Governing Board specified that some aspects of the standard setting process should be computerized to increase efficiency. WestEd, partnering with Measured Progress, was selected as the contractor for this work. They proposed a design including three operational sessions with 8 separate panels in each. Each session included two post-secondary areas—either college and one job training area or two job training areas, and two subjects—reading and mathematics. In addition to these 24 panels for the operational studies, a pilot study was conducted to include reading and mathematics panels for college preparedness and training programs in one occupational area—automotive master technicians. The entire standard setting process was computerized which introduced important efficiencies to the process and also provided data that led to additional enhancements to the process.[4]

Table 1 shows the pairings of post-secondary areas in sessions, and also gives panel identification codes used in Figures 1 and 2. Panels A and B are replicate panels with assignments of both panelists and items to each to be as equivalent as possible. All instructions were first provided in a general session to panelists in all eight panels, and the general sessions were followed by more specific instructions and training in panel groups under the direction of a process facilitator. For each operational session, four content facilitators provided content-based support, two for each post-secondary area. One content facilitator provided support to both replicate panels in each content area for each post-secondary area. Content facilitators worked with panelists to develop and modify the borderline performance descriptions to serve as the criterion for panelists' judgments. Facilitator training sessions were conducted in advance of the studies, and a facilitator handbook and a common set of instructional materials were developed to further promote consistency in instructions and procedures across the panels and sessions.

The operational sessions were to include 20 panelists for each post-secondary area in each subject—10 per replicate panel. Recruitment fell short of this requirement for one reading panel for college preparedness, for all automotive master technician panels, and for most of the pharmacy technician panels. In order to provide additional support to the job training panelists in the task of developing descriptions of the academic knowledge, skills, and abilities assessed in the NAEP items, the decision was made to recruit grade 12 teachers to supplement the math and reading job training panels for computer support specialist and HVAC technicians (the last two occupational areas in the series of studies) (Bay, 2012; WestEd and Measured Progress, 2012). Two high school teachers in each content area were recruited for each replicate panel in these two job training areas.

[4] A paper by Luz Bay, *Preparing Job Trainers to Describe Knowledge, Skills and Abilities Measured in an Academic Assessment*, presented in this session provides details regarding the development of descriptions of knowledge, skills, and abilities for each item and how that process was modified throughout the series of judgmental standard setting studies for grade 12 NAEP preparedness.

Table 1

JSS Sessions and Workshops[5]

JSS Session	Workshops	Panels	Panel Identifier
Pilot Study	College-Preparedness	Mathematics Panels A & B Reading Panels A & B	MCA & MCB RCA & RCB
	Automotive Master Technician	Mathematics Panels A & B Reading Panels A & B	MAA & MAB RAA & RAB
Operational Session 1	College-Preparedness	Mathematics Panels A & B Reading Panels A & B	MCA & MCB RCA & RCB
	Automotive Master Technician	Mathematics Panels A & B Reading Panels A & B	MAA & MAB RAA & RAB
Operational Session 2	Licensed Practical Nurse	Mathematics Panels A & B Reading Panels A & B	MLA & MLB RLA & RLB
	Pharmacy Technician	Mathematics Panels A & B Reading Panels A & B	MPA & MPB RPA & RPB
Operational Session 3	Computer Support Specialist	Mathematics Panels A & B Reading Panels A & B	MSA & MSB RSA & RSB
	HVAC Technician	Mathematics Panels A & B Reading Panels A & B	MHA & MHB RHA & RHB

Results of the Judgmental Standard Setting Studies

The pilot study and the three operational sessions resulted in cut scores that were quite different for replicate panels in several of the post-secondary areas, and they seemed unreasonable in light of other information available. Data in Tables 2 and 3 show cut scores on a pseudo-NAEP scale for mathematics and reading, respectively, for each pair of replicate panels within each post-secondary area. The data are transformed for reporting here such that they are based on the same transformation across the replicate panels and across the post-secondary areas. The NAEP score scale for mathematics ranges from 0 – 300 and for reading it is 0 – 500. A different scale was used for the two replicate panels and for the two post-secondary areas in each session of the judgmental standard setting studies to prevent comparisons across panels of results from round to round.

Note that the item pools were equally divided between panels A and B. Each panel rated half of the item pool plus at least two blocks of items that were common to the two panels. The item pools were divided to be statistically equivalent in terms of aspects such as item types, sub-domain content areas assessed, total possible score points, item difficulty, use of calculators in mathematics and provision of auxiliary materials. In reading, there was one passage with a lengthy poem and one passage that was a rental contract for which there was no counterpart to use in dividing the reading texts across the two replicate panels.

[5] This table is a modification of a table excerpted from page 39 of WestEd and Measured Progress (2011). *The National Assessment of Educational Progress Grade 12 Preparedness Research Project Judgmental Standard Setting (JSS) Studies: Process Report.* San Francisco.

While there is no reason to expect the level of minimal academic preparedness to be the same for training programs in different occupational areas, the expectation was that the judgments regarding minimal academic preparedness would not be statistically different for the two replicate panels in each occupation. For the most part, that was not the case—especially for reading. Mathematics cut scores in Table 2 for three pairs of replicate panels were not statistically different: Automotive, HVAC, and Pharmacy Technicians were similar for mathematics.

Table 2
Cut Scores on Common Pseudo-NAEP Scale for Mathematics Panels

Post-Secondary Area	Panel A	Panel B
College	425	413
Automotive Master Technicians	391	395
Computer Support Specialists	389	409
Heating, Ventilation, and Air Conditioning Technician	401	396
Licensed Practical Nurses	401	417
Pharmacy Technicians	398	400

Data reported in Table 3 show that for reading, the cut scores were statistically equivalent only for the replicate panels for HVAC. Scores for the other panels were considerably different.

Table 3
Cut Scores on Common Pseudo-NAEP Scale for Reading Panels

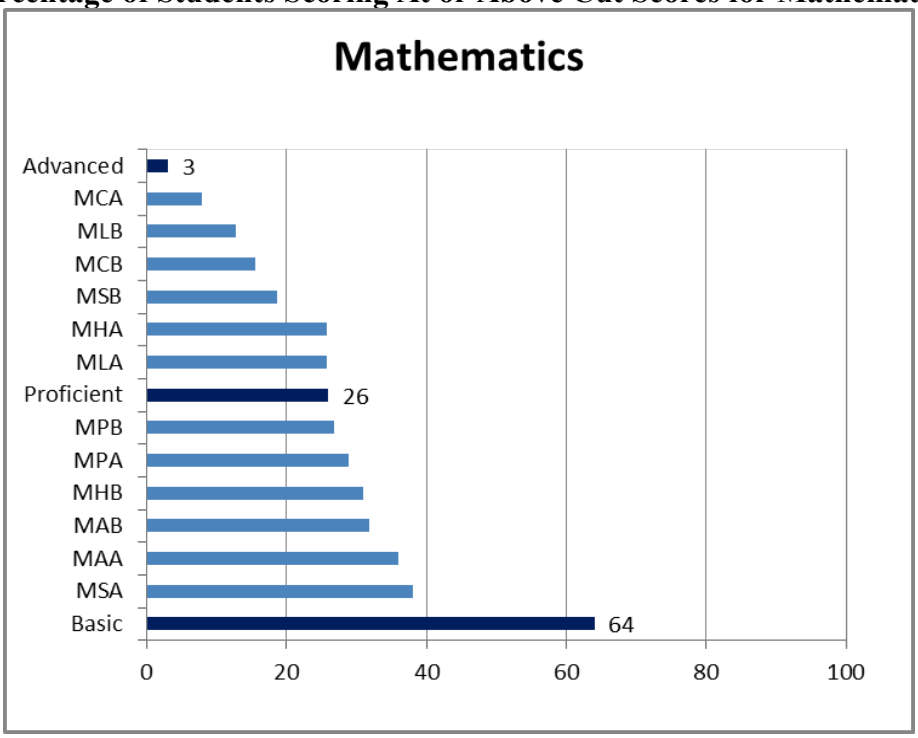
Post-Secondary Area	Panel A	Panel B
College	514	528
Automotive Master Technicians	532	518
Computer Support Specialists	518	531
Heating, Ventilation, and Air Conditioning Technician	513	516
Licensed Practical Nurses	531	512
Pharmacy Technicians	545	523

While the mathematics cut scores recommended for college course placement were higher than for the job training programs, this was not the case for reading. Figures 1 and 2 show the results reported as the percentages of 12^{th} grade students scoring at or above the cut scores set by each replicate panel. Data are also shown for the grade 12 achievement levels for the NAEP in mathematics and reading, respectively. The Proficient cut score for NAEP represents "*solid* academic performance; *competency* over *challenging* subject matter" (emphasis added). Data from the statistical linking studies of NAEP with SAT show that students who score at the Proficient level on the NAEP mathematics scale have about a .8 probability of scoring 500 on the math SAT (Moran, Oranje & Freund, 2011). The College Board has set 500 as the benchmark score to represent a .67 probability of a freshman year GPA of B- or higher (Wyatt, et al, 2010). It seems unlikely that a student would need to score much higher on NAEP to be placed in a

credit-bearing course that fulfills a general education requirement in mathematics than the score associated with an overall freshman year GPA of B-.

The data in Figure 1, however, show the cut scores for college course placement in mathematics (MCA and MCB) near the Advanced level for NAEP, and significantly higher than for most of the job training programs. This might lead to the conclusion that the minimal academic requirements to be prepared for college course placement in mathematics are higher than for job training programs. But, the cut scores representing the minimal level of academic preparedness in terms of performance on NAEP for the job training programs were also relatively high—generally near the NAEP Proficient cut score and higher for most of the panels in the occupational areas in the study. As previously mentioned, performance at the NAEP proficient cut score predicts a high probability of a college freshman year GPA of B- or higher.

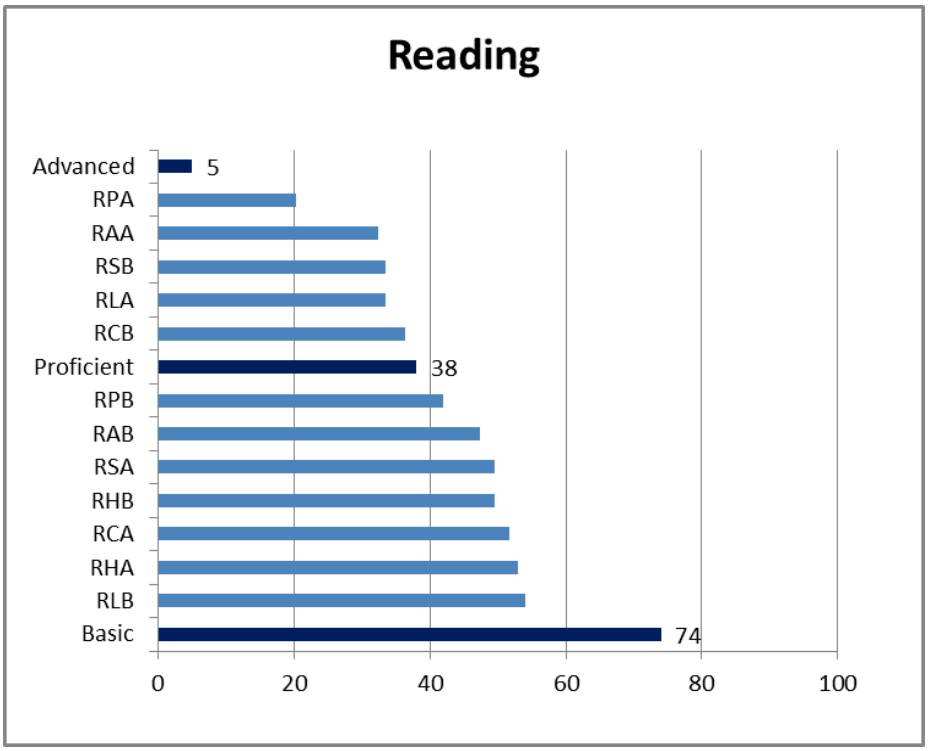
Figure 1
Percentage of Students Scoring At or Above Cut Scores for Mathematics



The statistical relationship between student performances on NAEP in reading and the SAT critical reading tests was weaker than that for mathematics (Moran, Oranje & Freund, 2011). The probability of scoring 500 or higher on the SAT critical reading was .5 for students scoring near the NAEP Proficient cut score. Again, a score of 500 on the SAT in critical reading predicts a freshman year GPA of B- or higher. In Figure 2, it appears that one panel set the cut score for minimal academic preparedness in reading for college course work at about the same score as the NAEP Proficient cut score, but most of the other panel set their score much lower. The five cut scores above the Proficient level were not very much higher on the NAEP scale. Based on the relative location of cut scores resulting from the judgmental standard setting studies, one might conclude that the minimal level of academic preparedness for college and careers is the same for

reading. But, the results were generally not replicated between the panels, and the relative relationships across the panels did not seem reasonable with respect to other information.

Figure 2
Percentage of Students Scoring At or Above Cut Scores for Reading



In addition to the comparisons to data from NAEP achievement levels and the SAT, Governing Board staff considered the information collected from panelists in advance of their participation in the judgmental studies, comments from the panelists during the process, and observational data collected during the panel studies. Those are covered next.

Other Findings from the Judgmental Standard Setting Studies

Knowledge, Skills and Abilities for Academic Preparedness

Prior to convening the panelists in St. Louis for the standard setting studies, webinars were conducted with panelists to provide a general orientation to the study, to the NAEP program, and to the Governing Board's preparedness research goals. At the end of the orientation webinars, panelists were instructed to sign on to a survey regarding NAEP content objectives for either reading or mathematics. Panelists were to evaluate each objective in terms of whether it is or is not required for a student to be at least minimally prepared for placement in a college-level course meeting the Governing Board's definitions or for entry in a job training program for one of the five occupations selected by the Governing Board for the research. The webinars were content-specific, and typically focused on job training programs in a specific occupational area, although panelists could join the webinar or play a recording of the webinar if their schedule did not permit joining the webinar for their specific post-secondary area.

As an example, of the relationship between panelists' judgments of NAEP mathematics content required for minimal preparedness, Table 4 shows responses for mathematics objectives rated by computer support and HVAC panelists prior to their participation in the panel studies. Note that the data collected from this process do not include responses from all panelists who participated in the studies.

Table 4
NAEP Mathematics Framework Objectives Rated as Required for Minimal Preparedness in Job Training Programs

Subcontent Area	Number of Objectives	% of Objectives in Each Subcontent Area Rated as Required by ≥50% of Panelists	
		Computer Support (15 respondents)	HVAC (13 respondents)
Number Properties and Operations	20	80%	70%
Measurement and Geometry	48	10	31
Data Analysis and Probability	32	25	9
Algebra	30	10	0

These responses indicate that preparedness in the area of number properties and operations was the most typical area of mathematical requirement for a majority of job training programs for both computer support specialists and HVAC technicians. Only about 15% of the total objectives for the grade 12 mathematics NAEP were allocated to this subcontent area, and the test specifications call for only 10% of the assessment items to cover this area. The mathematics framework objectives and test specifications appear to be relatively better aligned to the requirements of HVAC technicians than computer support specialists with respect to coverage of measurement and geometry. Measurement and geometry represents a larger proportion of the mathematics objectives (30%) and the largest proportion (37%) of test items specified for the assessment, and this area had a larger percentage of objectives rated as required for minimal preparedness by panelists for the HVAC mathematics study. The grade 12 mathematics test specifications called for algebra to have the highest proportion of items (35%), although a relatively low percentage of the 30 objectives in this area of mathematics was rated as required by a majority of panelists in either of these two sets of job training programs.

The survey information was collected to provide a basis for the content facilitators to draft descriptions of the knowledge, skills, and abilities that minimally prepared students would need to be eligible for the specific post-secondary activity—college course work or job training program. The reading content facilitators also requested that panelists provide a "typical" assignment for students entering their programs. These examples helped to better understand the reading skills required for the programs. Content facilitators drafted statements that served as

the first version of the borderline performance definitions (BPDs) for setting the preparedness cut score for the panels. The two replicate panels met together for their work on the BPDs because the same criteria were to be applied by both replicate panels for each post-secondary area. Panelists spent several hours working on the BPDs in the first two days of the session, and they had the opportunity to review and discuss the BPDs together prior to each round of bookmarking. Modifications to the BPDs could be made after rounds 1 and 2, but panelists were told in advance that only minor changes to the BPDs should be anticipated once ratings began. Panelists needed to agree that the BPDs were reasonable statements of what students should know and be able to do to be minimally prepared by the time panelists placed their first bookmarks.

When the content facilitators began working with panelists to evaluate the initial draft BPDs for the purpose of having panelists recommend needed modifications, a relatively high level of frustration among panelists was generally evident. Panelists had difficulty with the academic aspects of the NAEP frameworks. They were unsure of how the requirements of their training programs actually related to NAEP framework objectives, and they had difficulty in interpreting the framework objectives with respect to the mathematics or reading requirements of their training programs. The content facilitators were generally masterful in helping to translate the framework terminology so that panelists could relate the requirements for their students to the assessment criteria. Still, this was a difficult process. Panelists struggled with this task, and a huge effort was required to develop performance descriptions of the knowledge, skills, and abilities required for minimal preparedness.

Similarly, most panelists experienced difficulty in performing the task of describing the knowledge, skills, and abilities (KSAs) needed to correctly answer each item. This step is part of the training in preparation for bookmarking, but job training panelists found it nearly impossible to do this. The job trainers were not necessarily trained in mathematics or in reading; they were selected for their expertise in the content of the occupational training programs. The process evolved from asking panelists to write descriptions of the KSAs needed for each item to giving them descriptions of the KSAs for each item to discuss and record. (Bay, 2012)

Matching Student Populations Served by Job Training Programs with 12th Grade NAEP

Some panelists noted that the frameworks covered knowledge, skills, and abilities needed by their programs, but not at the level of the grade 12 assessments. Some panelists actually suggested that the grade 8 NAEP would have been a better match to the requirements of their programs. And, some information subsequently collected for additional research in this area indicate that a high school diploma or GED certificate is not required for as many as 42% of the occupational job training programs.

Further, the student populations served by the job training programs varied. While there was variability across panels within occupational areas with respect to whether students in the program were largely coming directly from high school or returning after an absence from formal education, the variability across occupational areas dominated. Variability ranged from a low percentage of students returning after an absence from education for job training as automotive master technicians (27%) and computer support specialists (34%) to a high percentage of this student population served by pharmacy technician (68%) and licensed

practical nurse (88%) job training programs. HVAC technician programs fell squarely in the middle with 50% of their programs serving students returning for training after an absence from education.

The Special Study of Irrelevant Items

The numerical results did not seem reasonable when compared to other data regarding preparedness with respect to performance on NAEP, and the results did not seem reasonable in light of the anecdotal evidence from panelists' comments and observations of their difficulties with the tasks. All of this information suggested that the cut scores resulting from the judgmental standard setting studies were not reasonable.

Other modifications to the design of studies had been implemented during the series of panel meetings. By the third operational session which involved the HVAC technicians and computer support specialists, the Governing Board staff determined that it was important to collect more research information regarding possible impacts on the results. The decision was made to seek more information about the items that panelists considered to be "irrelevant." This information would be helpful in evaluating the assessments and in evaluating the judgments of performances required to represent minimal academic preparedness.

Panelists had discussed "irrelevant" items from the start. Evidence of the perception of items as "irrelevant" was gathered in the pilot study, although it was not expected that these designations would be used for more than the initial development of the KSAs. Instructions for writing descriptions of KSAs required to respond correctly to each item were not clear in some panel groups, and some panelists apparently understood that an appropriate response would be "irrelevant," or "too hard." The fact that the standard setting process was fully computerized made it feasible to retain and review the descriptions panelists had entered as KSAs. This proved a valuable asset and led to clarification of instructions to panelists for the task (Bay, 2012). Among job trainers in automotive master technician programs, 58% of the KSAs written by one panelist were noted as "irrelevant," one panelist noted 42%, and one 35%. Other math panelists in these programs noted items as irrelevant, but generally fewer than 10% of the KSA descriptions referred to items as "irrelevant." For reading, only one panelist wrote descriptions including "irrelevant," and about 12% of that panelist's KSA descriptions referred to the irrelevance of the items.

The potential for "irrelevant" items had been considered in advance. The Governing Board's project director for this work had consulted with technical advisors having expertise in this area to discuss the potential that job trainers would lack the content knowledge required for 12th grade NAEP—especially for mathematics. The recommendation was to instruct panelists regarding the difference between "irrelevant" items and "items that are too difficult" for minimal preparedness. Following the pilot study, the Judgmental Standard Setting Technical Advisory Committee (JSS-TAC) recommended explicit instructions regarding placement of the bookmark with respect to irrelevant items. Panelists were instructed to place the bookmark before a string of irrelevant items if the next "relevant" item seemed too difficult to represent minimal preparedness. The issue had been confronted, at least to some extent, but the impact of irrelevant items on the outcomes of the studies was not clear. Thus, the decision was made to conduct the special study.

Item Maps

One change to the presentation of item maps for the bookmarking process had been made for these studies. For the Mapmark achievement levels-setting work (ACT, Inc. 2005, 2006, and 2009), item maps were organized with items mapped to a common score scale arranged within columns representing each subscale. For the preparedness standard setting studies, the Technical Advisors recommended that only color coding was necessary to identify items according to the subscale category. Since that was a change from the previous experiences for NAEP standard setting using a bookmark-based procedure, the special study design called for items to be mapped as originally planned, using the Mapmark format.

The primary consideration, however, for having the item maps more clearly differentiating the items by subscale was a result of the comments by some panelists indicating that entire subscales were irrelevant to the requirements for their training programs. Panelists were instructed to mark their cut score for each content area, one at a time, without consideration for either the location of the cut score previously set by the individual panelist or the median of individual panelists representing the cutscore for the replicate panel, and without consideration of the location of the cut score for other content areas in the special study. (Instructions to panelists for the special study are in Appendix A.) Panelists were given the option of marking an entire subscale as "irrelevant" in the special study. In fact, although some panelists marked most of the items in a subscale as irrelevant, none of the panelists marked an entire subscale as irrelevant.

Results of the Special Study

Panelists Judgments of Item Irrelevance in Relation to the Assessment Content

Both the job trainers for computer support specialists and HVAC technicians were engaged in the special study. Data are presented for only three panels in each subject for the heating, ventilation, and air conditioning technicians (HVAC), however, because the facilitators for two panels failed to implement the special study (mathematics panel A) or failed to implement it according to the instructions (reading panel A). The same item pool was rated by panelists in group A, no matter which occupation; likewise, the same item pool was rated by panelists in group B.

Data in Tables 5 – 7 show the percentages of mathematics items in each of the four subcontent areas rated as irrelevant by each panelist, and data in Tables 8 – 10 show the percentages for reading items. The count of "items" is actually the number of score points: multiple choice items (count as 1 each) plus the number of levels on constructed response items (score levels 1-3, for example).

Mathematics panelists in group A of the computer support specialists group rated *relatively* few items as irrelevant. Data in Table 5 show that no panelist in that group rated a majority of the items as irrelevant. By contrast, data in Table 6 show that each of the panelists in replicate panel B of the mathematics computer support specialists rated a majority of items as irrelevant, overall. Seven of the 10 panelists rated at least a majority of items in the number properties and operations area as irrelevant; all of the panelists rated a majority of items in measurement and geometry and in algebra as irrelevant. No panelist in that group rated a majority of items in the data analysis and probability set as irrelevant, however. The HVAC panelists in mathematics Panel B rated the same items as the Panel B computer support specialists. Five of the nine

HVAC panelists in the special study rated a majority of items in the number properties and operations subcontent area and in data analysis and probability as irrelevant; and six of the nine panelists rated a majority of items in algebra as irrelevant. A majority of items in the measurement and geometry subcontent area were rated as irrelevant by only three of the nine panelists.

Reading panelists had complained about the need for more relevant reading tasks—more passages of an "instructional manual" nature were recommended. (Please see tables 7-9 for reading data by panelists.) However, only one panelist (HVAC Panel B) rated a majority of items in the literary subcontent area as irrelevant. Otherwise, most panelists in Panel B of the computer support specialists rated none of the items as irrelevant while more panelists in Panel B of the HVAC technicians rated more items, especially those related to literary texts, as irrelevant.

Data in Tables 11 and 12 summarize this information for mathematics and reading, respectively. The data in those tables are reported as the percentage of items in each subcontent area rated as irrelevant in two ways: by at least one panelist and by at least 50% of the panelists in the group. This shows most clearly that the computer support specialists in math Panel B rated more items as irrelevant. For that panel, a total of 65 of the 111 items (59%) were rated as irrelevant by at least 50% of the panelists. In contrast, a majority of the HVAC panelists rating the same Panel B items considered only 41 of the 111 items (37%) to be irrelevant. Data in Table 12 show that 5 of the 29 literary items were rated as irrelevant by a majority of panelists in HVAC Panel B and 3 of the 68 informational items were rated as irrelevant by a majority of computer support panelists in Panel A. No reading items were rated as irrelevant by a majority of the computer support specialists in Panel B.

Additional analyses for the special study will be limited to mathematics panels. There is no basis for comparison for panel A across the training programs because there is only one set of ratings for panel A items; further, so few items were rated as irrelevant by the two B panels that further analysis did not seem warranted.

Table 5
Math Computer Support Specialists Panel A
Percent of "Items"[6] in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Number Properties and Operations (9)[7]	0	11	0	0	22	22	0	0	11	11
Measurement and Geometry (47)	16	7	16	13	18	13	7	7	22	16
Data Analysis and Probability (21)	5	10	10	14	5	10	29	10	19	10
Algebra (36)	14	14	22	14	17	11	11	17	11	17
Total (113)	12%	10%	15%	12%	15%	12%	12%	10%	17%	14%

Table 6
Math Computer Support Specialists Panel B
Percent of "Items" in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Number Properties and Operations (12)	42	50	42	50	75	58	50	42	67	58
Measurement and Geometry (40)	60	60	88	73	93	65	60	65	90	80
Data Analysis and Probability (30)	27	33	37	40	43	37	47	37	40	40
Algebra (29)	59	59	69	55	62	52	55	62	55	59
Total (111)	59%	51%	64%	57%	69%	53%	54%	54%	65%	61%

Table 7
Math Heating, Ventilation, and Air Conditioning Technicians Panel B
Percent of "Items" in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9
Number Properties and Operations (12)	42	83	50	8	58	50	33	0	83
Measurement and Geometry (40)	25	85	20	10	20	58	13	5	70
Data Analysis and Probability (30)	17	83	60	10	53	53	33	17	80
Algebra (29)	10	79	62	10	48	69	66	0	55
Total (111)	21%	83%	45%	10%	41%	59%	34%	6%	70%

Table 8
Reading Computer Support Specialists Panel A
Percent of "Items" in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Literary (28)	4	7	11	7	0	0	0	0	0	32
Informational (68)	10	15	2	6	4	16	16	16	4	34
Total (96)	8%	13%	4%	6%	3%	12%	12%	12%%	3%	33%

Table 9

Reading Computer Support Specialists Panel B
Percent of "Items" in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Literary (29)	0	0	0	0	0	2	0	0	0	2
Informational (71)	0	0	0	0	4	3	0	0	0	2
Total (100)	0	0	0	0	4%	5%	0	0	0	4%

Table 10

Reading Heating, Ventilation, and Air Conditioning Technicians Panel B
Percent of "Items" in Each Subscale Marked Irrelevant by Each Panelist

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Literary (29)	0	31	24	34	24	17	24	17	59	34
Informational (71)	0	15	20	0	6	8	8	7	7	7
Total (100)	0%	20%	21%	10%	11%	11%	13%	10%	22%	15%

Table11
Summary of Mathematics Items Rated Irrelevant

Criterion	Content Area	% Computer Support Panel A	% Computer Support Panel B	% HVAC Panel B
Rated as Irrelevant by at Least One Panelist	Numbers and Operations (A=9, B=12)	56	75	92
	Measurement and Geometry (A=47, B=40)	100	100	95
	Data Analysis and Probability (A=21, B=30)	86	67	100
	Algebra (A=36, B=29)	94	86	93
Rated as Irrelevant by at Least Half of the Panelists	Numbers and Operations (A=9, B=12)	11	50	50
	Measurement and Geometry (A=47, B=40)	79	78	20
	Data Analysis and Probability (A=21, B=30)	17	33	40
	Algebra (A=36, B=29)	14	62	52

Table 12
Summary of Reading Items Rated Irrelevant

Criterion	Content Area	% Computer Support Panel A	% Computer Support Panel B	% HVAC Panel B
Rated as Irrelevant by at Least One Panelist	Literary (A=28, B=29)	86	32	14
	Informational (A=68, B=71)	46	51	10
Rated as Irrelevant by at Least Half of the Panelists	Literary (A=28, B=29)	17	0	0
	Informational (A=68, B=71)	0	4	0

Recall from Table 4 that when evaluating the objectives in each subcontent area prior to participating in the actual JSS study, a majority of respondents in the HVAC technicians and computer support specialists groups rated number properties and operations as the area with the highest proportion of objectives required for minimal preparedness in their job training programs. After working with the items in the JSS process, however, a majority of these job trainers rated 50% of the number properties and operations items in the Panel B pool as irrelevant. The computer support specialists had rated only about 10% of mathematics objectives in measurement and geometry as representing required KSAs for minimal preparedness, and a majority of these panelists rated large percentages of the items in this subcontent area as irrelevant. None of the algebra objectives were considered as requirements for minimal preparedness to enter HVAC training programs by a majority of respondents prior to the JSS studies; however, only 52% of the algebra items were rated as irrelevant by a majority of panelists in the study.

While there is some consonance between the mathematics areas considered important to minimal preparedness and the relevance of items developed to measure those objectives, the agreements are not strong. It is not clear whether the judgments of the panelists changed after gaining a clearer understanding of how the objectives were operationalized in the form of test questions, or if there was just a misunderstanding of the objectives at the start. In any case, the bookmarks were placed relative to borderline performance descriptions that were initially based on those ratings of objectives, and this analysis provides evidence that substantial numbers of items that were considered irrelevant were in areas that were originally judged to be important to minimal preparedness.

Panelists Judgments of Item Irrelevance in Relation to Preparedness Cutscores

For the special study, items were arranged on item maps according to the subcontent area. Panelists were instructed to place their bookmark for items in each subcontent area, using the same criteria and procedures used through the study. But, this time, they had to mark their cut scores by hand on a paper copy of the item maps rather than via the computerized standard setting software. They could start with any subcontent area. Table 13 shows data for the minimum, maximum and median cut scores for each subcontent area in mathematics and Table 14 shows the data for reading.

As the data in Tables 13 and 14 demonstrate, panelists generally placed the bookmark in the subcontent areas such that the median cutscore for the panel group was not very different from the cut score set in Round 3 of the bookmarking process. (Cut score data are reported here on a pseudo-NAEP scale.) To the extent that there were differences of more than one or two points, the medians derived from the weighted cutscores for each separate subcontent area were higher than the median cut score set in Round 3 for the group. That was especially true for the HVAC job trainers in mathematics Panel B who set the cut score 16 points higher in the special study than in Round 3.When asked whether a cut score computed from the separate subcontent cut scores was likely to be the same, higher, or lower than the group cut score for Round 3, 78% of the respondents said they expected it to be about the same.

Table 13

Cut Score Data for Mathematics Panels Using Items Mapped by Subcontent Area

Mathematics Subcontent Area	Statistics	Computer Support Panel A	Computer Support Panel B	HVAC Panel B
Numbers and Operations	Minimum	384	399	350
	Maximum	448	437	503
	Median	395	409	434
Measurement and Geometry	Minimum	390	339	397
	Maximum	426	427	525
	Median	392	408	427
Data Analysis and Probability	Minimum	380	399	376
	Maximum	424	429	447
	Median	390	413	399
Algebra	Minimum	364	393	376
	Maximum	422	425	451
	Median	390	406	401
Weighted Average of Medians		391	409	412
Round 3 Cut Score		389	409	396

Note: Pseudo-NAEP cut scores are reported in the table. The NAEP scale for mathematics is 0 – 300.

Table 14

Cut Score Data for Reading Panels Using Items Mapped by Subcontent Area

Content Area	Statistics	Computer Support Panel A	Computer Support Panel B	HVAC Panel A	HVAC Panel B
Literary	Minimum	511	502	710	726
	Maximum	532	532	516	546
	Median	515	529	512	513
Informational	Minimum	512	513	502	512
	Maximum	582	548	523	530
	Median	522	531	517	516
Weighted Average of Medians		515	520	530	515
Round 3 Cut Score		513	516	531	516

Note: Pseudo-NAEP cut scores are reported in the table. The NAEP scale for reading is 0 – 500.

Panelists were generally positive in their evaluation of the procedure of placing bookmarks for items in each separate subcontent area: 85% responded that this procedure helped them to represent minimal preparedness requirements on the NAEP items more accurately. Nearly three-quarters of the respondents (72%) reported that they at least somewhat agreed that it was easier to deal with the issue of irrelevant items when placing their bookmarks for items in separate subcontent areas.

Item Maps

Two item maps marked up to show items judged to be irrelevant by at least half of the panelists in each group are presented for illustrative purposes[8]. These two maps in Figures 3 and 4 were used by HVAC technicians and computer support specialists job trainers, respectively, in the mathematics Panel B groups. Items rated as irrelevant by at least half of the panelists in the group are blackened on the item maps. At first glance, it appears that the two groups rated the same items differently. A majority of the job trainers for HVAC technicians rated many fewer items as irrelevant than was the case for the job trainers for computer support specialists. A majority of the HVAC group rated none of the items below the group cut score as irrelevant, although a majority of the computer support group rated a few of those as irrelevant. A majority of computer support specialists found most of the items in the measurement and geometry subcontent area to be irrelevant, but HVAC panelists found few to be irrelevant. Table 11 reported the percentages of items in each subcontent area rated as irrelevant by these panelists.

Interestingly, there was rather high agreement between these two Panel B groups on a sizable number of items rated as irrelevant. A majority of panelists in both groups rated exactly the same six items as irrelevant in number properties and operations. A majority of panelists in both groups rated the same nine items in measurement and geometry as irrelevant. Of course, a majority of computer support specialists also rated 22 other items in the measurement and geometry area as irrelevant. Eight of the data analysis and probability items rated as irrelevant by a majority of panelists in the HVAC group were rated as irrelevant by a majority of computer support panelists; and there was only one additional item in that area rated as irrelevant by at least half of the computer support specialists. Finally, 13 of the 15 algebra items rated as irrelevant by a majority of HVAC panelists were also rated as irrelevant by the computer support panelists. So, there was actually fairly high agreement by a majority of panelists in each group on the "irrelevance" of one-third of the items in the grade 12 NAEP mathematics item pool.

[8] With so few items marked as irrelevant by a majority of reading panelists, the specially coded item maps did not seem worthwhile.

Figure 3

2009 NAEP Mathematics
Heating, Ventilation, and Air Conditioning Technicians

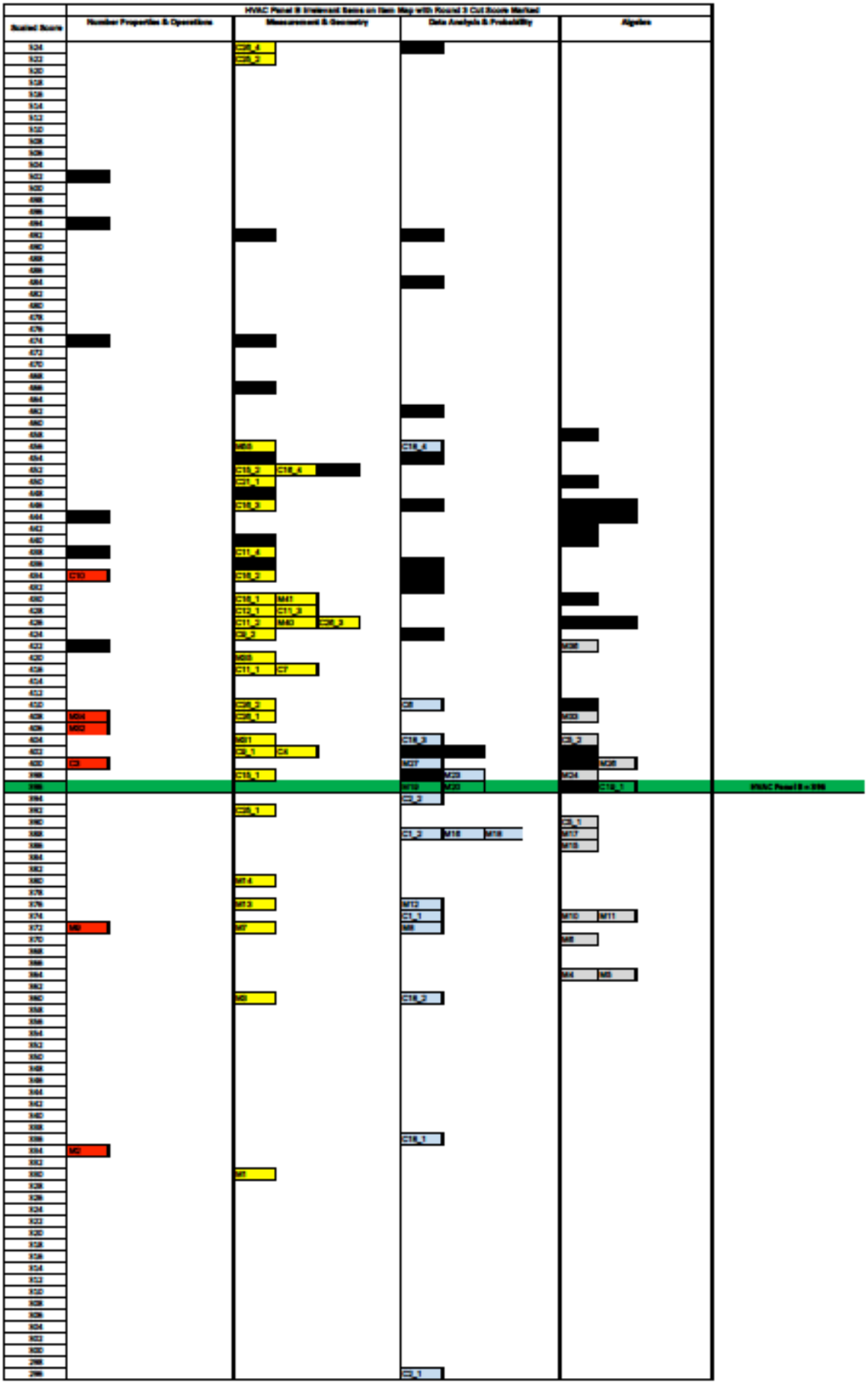
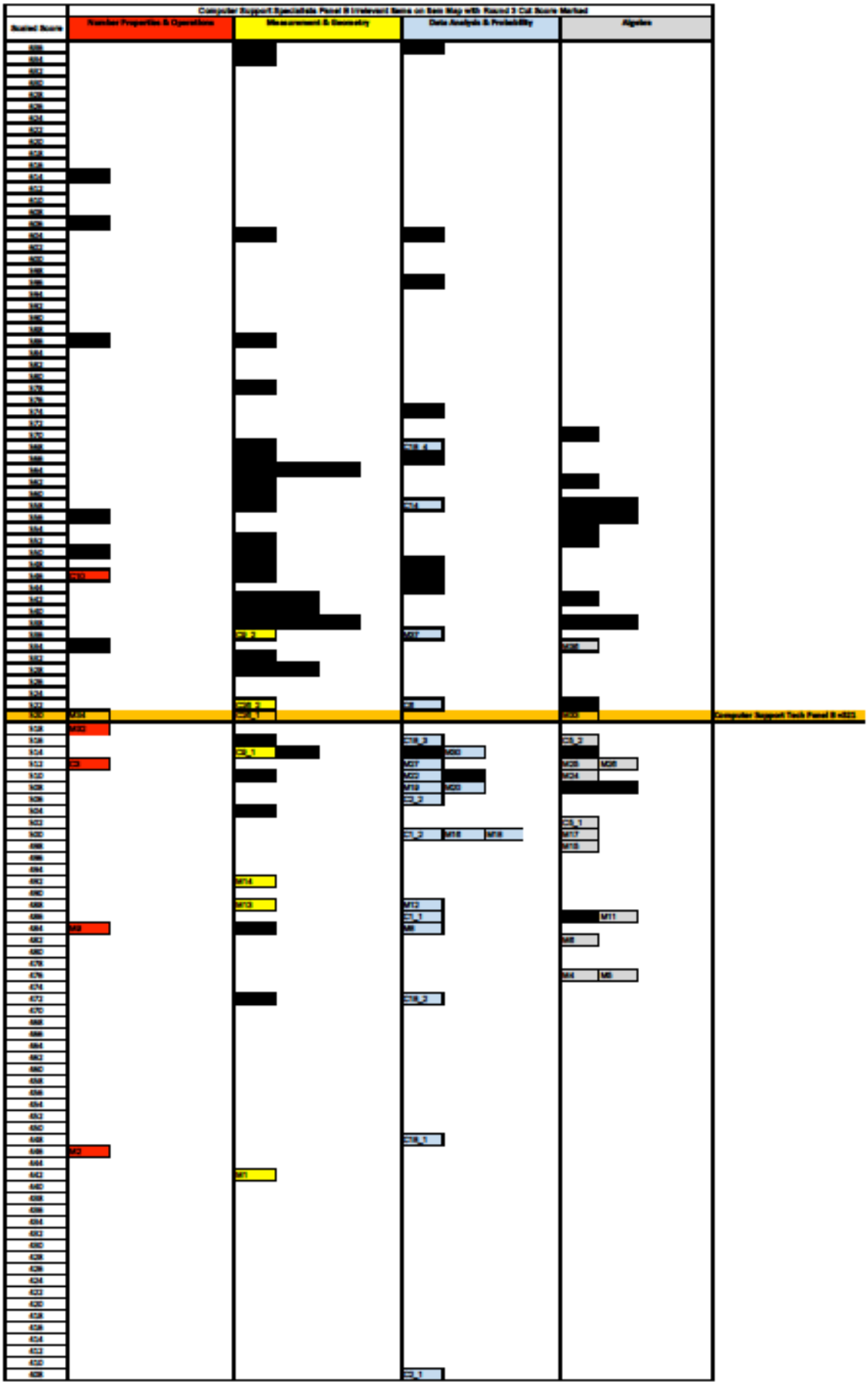


Figure 4

2009 NAEP Mathematics
Computer Support Specialists



Summary

Panelists in the judgmental standard setting studies for grade 12 NAEP preparedness research faced a greater challenge than had been anticipated. When planning the studies, consideration was given to the potential for a lack of content knowledge on the part of job trainers who were selected for their expertise in the occupational areas, rather than in the content of reading or mathematics. The assumption was that the items might be too difficult, especially mathematics items, for panelists who are not required to use the range of knowledge and skills assessed in NAEP on a regular basis. Similarly, while no problems in reading, per se, were anticipated, there was some concern about the ability of panelists to judge the skills required of job training students to perform the literary-based tasks in the reading NAEP. There was an understanding that some of the items and content on grade 12 NAEP in both reading and mathematics would not be necessary for minimal academic preparedness in the job training occupations. The extent to which items were considered irrelevant had not been fully anticipated, however.

The task of placing a bookmark to represent minimal academic preparedness for entry in the job training programs was far more complex than anticipated, due to the perception of items as irrelevant. The distinction between items that were simply too difficult to represent minimal preparedness and items that were truly unrelated to the job training requirements seemed to be quite subtle and difficult to discern.

Additional research is underway to better understand the knowledge, skills, and abilities needed for job training programs in the five occupational areas studied for the 2009 grade 12 NAEP preparedness research. The goal is to evaluate course materials for both entry level and exit level courses in the programs, starting with the entry level courses. Materials such as the course syllabus, text books, assignments, tests, and so forth are being collected and evaluated to document the KSAs that are considered prerequisite for entering students, those that are reviewed in the course, and those for which there is no expectation of any prior learning. The course materials are being coded relative to the objectives of the NAEP framework in mathematics and in reading, supplemented by vocational and technical career standards. Additional KSAs are added when the review team determines the need to do so. Those KSAs will then be compared to the descriptions of minimal academic preparedness developed for each job training area and to items in the NAEP item pool near the cut scores recommended for the job training area by panelists in the JSS studies. The goal is to develop a clear description of the knowledge, skills, and abilities needed for the job training programs and to determine the extent to which the grade 12 reading and mathematics assessments measure them. The Governing Board plans an extensive review of the frameworks and assessments after this information is collected.

Although the results of these studies did not lead to points to be reported on the NAEP scale to serve as indicators of academic preparedness in the job training programs, valuable information was generated through the studies. The results of these studies point to important differences between the knowledge, skills, and abilities needed for entry in these job training programs and the knowledge, skills, and abilities needed for college course work. In addition, the findings of these studies magnify the necessity of a strong alignment between criteria for judgmental standard setting and the assessment for which the criteria are applied. The recommendations for

"preparedness" modifications to the NAEP grade 12 frameworks (Achieve, 2006a; 2006b) focused on preparedness for "high trajectory" jobs. The criteria for selection of job training program occupations for these studies were not exclusively "high trajectory jobs." Instead, they focused on occupational areas that do require at least three months of post-secondary training, but do not require a bachelor's degree. This is an important distinction in that it does not equate college and career preparedness in terms of the necessary educational credentials. The clear indication of the findings of these studies is that there are important and significant differences in the academic requirements associated with preparedness for college coursework and job training programs. This is truly an area for which *much more research is needed*.

References

Achieve, Inc. (2006a). *Recommendations to the National Assessment Governing Board on Aligning 12th Grade NAEP with College, Workplace, and Military Expectations: Mathematics.* Washington, DC: Author.

Achieve, Inc. (2006b*). Recommendations to the National Assessment Governing Board on Aligning 12th Grade NAEP with College and Workplace Expectations: Reading*. Washington, DC: Author.

ACT, Inc. (2005). *Developing Achievement Levels on the 2005 National Assessment of Educational Progress in Grade Twelve Mathematics: Process Report.* Iowa City, IA: Author.

ACT, Inc. (2007). *Developing Achievement Levels on the 2006 National Assessment of Educational Progress in Grade Twelve Economics: Process Report.* Iowa City, IA: Author.

ACT, Inc. (2010). *Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Grades 4, 8, and 12 Science: Process Report*. Iowa City, IA: Author.

Bay, L. (2012). *Preparing Job Trainers to Describe Knowledge, Skills and Abilities Measured in an Academic Assessment.* Paper presented in the *Setting Academic Preparedness Standards for Job Training Programs: Are We Prepared?* session at the annual meetings of the National Council on Measurement in Education, April 14, 2012, Vancouver, British Columbia, Canada.

Wyatt, J., Kobrin, J., Wiley, A., Camara, W. & Proestler, N. (2010). *Development of a College Readiness Index and its Relationship to Secondary and Post-Secondary School Performance.* New York: The College Board.

Moran, R., Oranje, A., & Freund, D. (2011). *NAEP 12th Grade Preparedness Research: Establishing a Statistical Relationship between NAEP and SAT.*® Technical Memo to The National Assessment Governing Board, Committee on Standards, Design and Methodology, December 2011.

National Assessment Governing Board (2009). *Making New Links: 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*. Washington, DC: Author.

National Assessment Governing Board (2008a). *Mathematics Framework for the 2009 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board (2008b). *Reading Framework for the 2009 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board (2004). *12th Grade Student Achievement in America: A New Vision for NAEP. A Report to The National Assessment Governing Board.* Washington, DC: Author.

WestEd & Measured Progress. (2011). *National Assessment of Educational Progress Grade 12 Preparedness Research Project Judgmental Standard Setting (JSS) Studies: Process Report.* San Francisco, CA.