

The Modified Briefing Book Standard Setting Method: Using Validity Data as a Basis for Setting Cut Scores

Julie A. Miles, Jennifer N. Beimers, and Walter D. Way
Pearson Assessment & Information

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 30, 2010 – May 3, 2010 in Denver, CO.



*Using assessment
and research to
promote learning*

Abstract

The dominant standard setting approaches in large-scale educational assessment are squarely focused on test content. Commonly used methods such as Bookmark and modified Angoff require judgments based on an analysis of test items and content-based performance level descriptors. Similarly, portfolio-based methods, such as Body of Work, involve evaluating specific examples of examinee performance. For statewide standards-based tests, a content focus is entirely appropriate; the established content-based standard setting methods provide the process documentation needed to establish and defend cut scores for high-stakes statewide assessments. In this paper, we describe a standard setting approach that was developed for a high school end-of-course Algebra II exam used by a consortium of 15 states. A major purpose of this exam is to serve as a measure of readiness for instruction in college credit-bearing courses. To support this use of the exam, it was imperative to develop a standard setting process that would be informed by data describing relationships between test performance and external criteria. This paper focuses on two aspects of the modified briefing book standard setting process developed to meet this need: 1) the validity research conducted to support the standard setting; and 2) the standard setting itself, through which the validity research and associated pertinent information was organized and presented to the panelists, and resulting process through which these data were used to elicit cut score judgments.

Keywords: standard setting, college readiness, validity

The Modified Briefing Book Standard Setting Method: Using Validity Data as a Basis for Setting Cut Scores

The dominant standard setting approaches in large-scale educational assessment are squarely focused on test content. Commonly used methods such as Bookmark and modified Angoff require judgments based on an analysis of test items and content-based performance level descriptors. Similarly, portfolio-based methods, such as Body of Work, involve evaluating specific examples of examinee performance. For statewide standards-based tests, a content focus is entirely appropriate; the established content-based standard setting methods provide the process documentation needed to establish and defend cut scores for high-stakes statewide assessments.

One drawback of these judgment-based standard setting methods is that cut-scores are seldom set with external references to criteria that may be related to test performance. For example, because proficiency levels are set by each state in reference to its own content standards and policy considerations, the level of test performance necessary to reach “proficient” or “advanced” differs markedly across states.

In this paper, we describe a standard setting approach that was developed for a high school end-of-course Algebra II exam used by a consortium of 15 states. A major purpose of this exam is to serve as a measure of readiness for instruction in college credit-bearing courses. To support this use of the exam, it was imperative to develop a standard setting process that would be informed by data describing relationships between test performance and external criteria. This paper focuses on two aspects of the standard setting process approach developed to meet this need: 1) the validity research conducted to support the standard setting; and 2) the standard setting itself, through which the validity research and associated pertinent information was organized and presented to the panelists, and resulting process through which these data were used to elicit cut score judgments.

Background on the ADP Algebra II Exam

The American Diploma Project (ADP) was initiated by Achieve, Inc. to ensure that all students graduate from high school prepared to face the challenges of work and college. The ADP Network now includes 35 states—responsible for educating nearly 85 percent of all U.S. public school students.

In the fall of 2005 with support from Achieve, nine American Diploma Project Network states—Arkansas, Indiana, Kentucky, Maryland, Massachusetts, New Jersey, Ohio, Pennsylvania and Rhode Island—came together to develop specifications for a common end-of-course exam in Algebra II. Six additional states—Arizona, Florida, Hawaii, Minnesota, North Carolina and Washington—have since joined the ADP Assessment Consortium, bringing the total number of participating states to fifteen.

In March 2007, the states awarded the contract to develop and administer the ADP Algebra II End-of-Course Exam to Pearson. Field testing of Algebra II End-of-Course Exam items was conducted in October 2007 and February 2008. The Algebra II exam was first administered in spring 2008 to nearly 100,000 students across the participating states (Achieve, 2008). However, Achieve and the participating states made a deliberate decision not to set cut scores on the basis of this first exam administration. Rather, a series of validity studies were designed and conducted to inform the standard setting that was planned to take place after the spring 2009 administration.

At the inception of the program, the Algebra II exam consisted of 46 multiple-choice items, seven short-answer constructed response items worth two points each, and four extended-response constructed response items worth four points each, with a total possible raw score of 76. Following feedback from the first operational administration, the exam was shortened to consist of 46 multiple-choice items, six short-answer constructed response items worth two points each, and three extended-response constructed response items worth four points each, with a total possible raw score of 70. The exam is divided into a calculator and non-calculator portion, and measures the following content strands: Operations on numbers and expressions, equations and inequalities, polynomial and rational functions, exponential functions, and function operations and inverses.

The American Diploma Project Algebra II End-of-Course Exam is designed to serve 3 critical goals:

- to improve high school Algebra II curriculum and instruction;
- to serve as an indicator of readiness¹ for first-year college credit-bearing courses; and
- to provide a common measure of student performance across states over time

Standard Setting Methodology

The method used to set standards for the Algebra II exam is referred to as the Modified Briefing Book method, due to similarities with a method suggested by Haertel (2002; 2008). In this method, the standard setting process is informed by a “briefing book”, in which a compendium of relevant information to inform a standard setting is compiled and made available to the participants in the standard setting process. Our approach differed from Haertel’s in that we did not develop overt discussions of, say, 10 alternative cut scores. However, we did bring together a variety of policy background, research data, test content information, and data about student performance in a comprehensive and focused fashion designed to structure participants’ input and policymakers’ deliberations. Thus, in both intent and outcome, our approach seems very consistent with Haertel’s concept of standard setting as a participatory process.

¹ For the purposes of the ADP Algebra II assessment and setting of cut scores, “college readiness” was operationally defined as students who would likely earn a ‘B or better’ in their first college-level credit-bearing course without prior remediation.

Briefing Book Contents

The heart of the briefing book used at the Algebra II standard setting was a series of validity studies carried out with the exam in fall 2008 and spring 2009. The validity studies included in the briefing book focused on the use of the exam as an indicator of readiness for first-year college-level credit-bearing courses². These included:

1. Concurrent studies—Student scores from the Spring 2008 ADP administration were matched to student scores to other state and national assessments to establish relationships, including those with existing measures of college readiness.
2. Cross-sectional studies—The ADP Algebra II Exam was administered to students at the beginning of the semester of their college mathematics course and compared to their final grade in the course to determine how well a student’s performance on the exam predicts his/her performance in the college math course.
3. Judgment studies—Feedback was gathered from 133 college professors who teach College Algebra and Pre-Calculus courses regarding the relevance of the ADP Algebra II Exam standards to their courses. In addition, these participants drafted performance level descriptors, and evaluated the exam with respect to their expectations of what students need to have previously mastered in order to successfully³ learn the material covered in their Algebra or Pre-Calculus courses.

In addition to these three major studies, content experts working with the Algebra II exam conducted an exercise through which they mapped items to the performance level descriptors (PLDs) developed during the judgment studies. In this exercise the content experts evaluated the content of each Algebra II test item and assigned it to the most appropriate performance level based on the PLDs. The resulting pattern of assigned proficiency levels was then examined for potential cut score locations to provide an additional piece of context for the standard setting process.

Also included in the briefing book, to provide a wider context to the validity studies, were results from Achieve-conducted content-based studies that describe the variability of Algebra II standards within the United States, the limited focus on Algebra II content found on existing college admissions and placement tests, the Algebra II content included in a sample of College Algebra and Pre-calculus college courses, and a comparison of the Algebra II exam standards to international standards.

² The other two goals of the program will be fulfilled through more long-term studies, such as longitudinal studies following a cohort of students from high school, where they took the ADP Algebra II Exam, into college and analyzing their exam grade, courses taken in high school, and performance in college.

³ “Successfully” was defined as earning a B or better as a final course grade without prior remediation.

The complete ADP Algebra II Standard Setting Briefing Book is available for review and download on the [PearsonAccess website](#).

Summary of Validity Studies

Concurrent Studies. The concurrent studies examined relationships between student scores on the spring 2008 Algebra II exam and scores on national exams and state exams. The concurrent validity studies provide indirect information for considering the use of the ADP Algebra II exam to assess college readiness. For example, the SAT and ACT already provide implicit and explicit benchmarks of college readiness. The PSAT is used as a qualifying exam for the National Merit Scholarship program, and the state tests classify students into “Proficient” and “Advanced” performance categories which could be useful in providing additional context to the state representatives attending the standard setting.

The national exam data were based on 7,277 students taking the ACT mathematics subtest, 619 students taking the SAT mathematics section, and 954 students taking the PSAT mathematics section. In addition, relationships between ADP Algebra II scores and statewide mathematics exam scores for a total of 9,547 students across six states were summarized. The analyses undertaken for the concurrent studies included compilation of univariate statistics, correlations, regression analyses, and applications of equipercentile equating methods to establish concordance relationships.

For the ACT and SAT studies, logistic regression was conducted in which a specific level of achievement (0=below, 1=at or above) was regressed on Algebra II exam scores. Results were compiled in expectancy tables that represented the probability of achieving a certain ACT or SAT score or better based on Algebra II exam performance. Additionally, by creating a concordance table of ACT and ADP test scores for the matched student samples, we were able to map the ACT ‘college-ready’ score of 22 to an ADP exam score of 25 and 26 out of 76 points. Another way of analyzing the data was using linear regression to see what score on the ADP exam predicted an ACT score of 22. The results indicated that an ADP score of 31 predicted an ACT score of 22. This led us to the final analyses using a logistic regression model to determine what score on the ADP exam had at least a 65% probability of being associated with a 22 on ACT. This analyses resulted in a score of 32 on the ADP exam have a 65% probability of earning an ACT score of 22.

The executive summary for the concurrent studies presented in the briefing book is shown in Appendix A⁴. The summary presents highlights of the studies: background, participants, methods used, various results, and considerations of the limitations of the data.

Cross-Sectional Studies. The cross-sectional studies were based on 3,132 college students from 31 different institutions taking college algebra or pre-calculus courses. The

⁴ Analyses for SAT scores were also done, however due to the lack of generalizability of the sample the results were presented with caveats.

students were administered the Algebra II exam at the beginning of the semester. Students were split fairly evenly between different types of institutions (i.e., two-year colleges, four-year colleges defined as “typical” based on admission rates, and four-year colleges defined as “selective” based on admission rates). When analyzing the college student performance, two methods were used to draw conclusions about how well college students would perform on this exam.

In the first analyses, a regression method was used to determine what score on the ADP exam had at least a 65% probability of being associated with final course grade of “B or better” which is our working definition of ‘college-ready’. This method yielded a score range between 32 to 38 depending on the course-type and institution type included in the analyses. The second method was to look at the data using contrasting groups analyses where we looked for the point of intersection between two distributions of students. Since we’ve defined college-ready as ‘B or better’ we contrasted all students with a final course grade of C/D/F with students earning B/A. This method indicates that an ADP score of 25 separates the two groups.

The executive summary for the cross-sectional studies presented in the briefing book is shown in Appendix B. The summary presents highlights of the studies: background, participants, methods used, various results, and considerations of the limitations of the data.

Judgment Studies. For the judgment studies, three one-day meetings were held in the spring of 2009 with 133 professors representing 79 institutions and 20 states. Professors teaching both college algebra and pre-calculus for community colleges, four-year institutions with a typical admittance rate and four-year institutions with a more selective admittance rate were represented. Participants were presented with three major tasks:

1. a standards relevance survey, in which each Algebra II content benchmark was rated as to its relevance in preparing successful students for the first college level credit-bearing math course;
2. defining PLDs, through which participants described what knowledge, skills, and abilities students in each of the performance levels would be able to demonstrate on the first day of their college-level math course; and
3. item level judgments similar to those rendered in a traditional item-mapping standard setting through which an estimate of the number of points a “successful” student would earn on the first day of class was tallied for each participant.

Two outcomes of the judgment studies assisted us in connecting the ADP exam to the college-ready expectations of faculty who teach College Algebra and Pre-Calculus. The first outcome was a list of core competencies and descriptions of what students should know in each performance level that we will be recommending cuts for during this meeting. The PLDs were drafted by the college faculty at the Judgment Studies and revised and approved by Pearson, Achieve, and the Assessment Consortium. The second

outcome, was their recommendation on where the Prepared cut score should be located using the content of the exam as their guide. Using this content-based evaluation, the results indicated that college faculty expected a score between 23 and 38 depending on course type and institution type.

The executive summary for the judgment studies presented in the briefing book is shown in Appendix C. The summary presents highlights of the studies: background, participants, methods used, various results, and considerations of the limitations of the data.

Additional Studies. A different lens to view the content through is to use the PLDs as the basis for evaluating individual items; therefore, in addition to the finding from these three major series of studies presented in the briefing book, content experts working with the Algebra II exam conducted an exercise through which they mapped individual test items to the PLDs developed during the judgment studies. In this exercise the content experts evaluated the content of each Algebra II test item and assigned it to the most appropriate performance level based on the PLDs. The resulting pattern of assigned proficiency levels was then examined for potential cut score locations (yielded possible cut scores of 32 for Prepared and 45 for Well-Prepared) to provide an additional piece of context for the standard setting process.

To summarize all the data from the various studies and aide standard setting panelists in their review of the results, a crosswalk spreadsheet was constructed that showed the results of the studies in a side-by-side fashion. The leftmost column contained all possible raw score points and subsequent columns were dedicated to each of the studies. The results of each study (potential cut score regions) were then identified in the cells that corresponded to the appropriate raw scores. Those studies that were considered most significant to the standard setting process, based on the limitations and generalizability of the data, were highlighted. The crosswalk presented in the briefing book is shown in Appendix D.

The crosswalk illuminated the converging of data which illustrated the rigor of the ADP Algebra II exam. Both the empirically-based and judgment studies seemed to support placing the Prepared cut score within a range of scores between 24 and 38.

Standard Setting Meeting

Standards were set on the American Diploma Project (ADP) Algebra II End-of-Course Exam during a two-day meeting in July 2009. As mentioned previously, the results of the above described studies were summarized and organized into a “briefing book” that was mailed to standard setting participants prior to the standard setting. Twenty-seven panelists including representatives from all fifteen ADP consortium states’ departments of education and higher education representatives determined the recommended boundaries between the three performance levels being reported on: needs preparation, prepared, and well prepared. These performance levels describe a student’s

preparedness to succeed in a first-year credit-bearing college mathematics course without remediation.

The standard setting meeting was facilitated by Pearson, the developer/owner of the ADP Algebra II end-of-course exam. At the request of the consortium states, Achieve staff attended the standard setting as an observer and carried the final responsibility for setting the cut scores for the exam based on the recommendations from the standard setting panel along with input provided by the assessment consortium’s coordination and direction team (CDT).

Who Were the Panelists?

The standard setting panel included 15 state department of education representatives (one representative for each of the 15 exam consortium states) and 12 mathematics and higher education representatives. Panelists were separated into six tables with state and higher education representatives mixed within tables. The group was primarily male (63%) and predominately white (70%). Table 1 shows the breakdown of various demographics of the panelists.

Table 1. Demographics of the Panelists

Gender		Ethnicity	
Male	63%	American Indian	0%
Female	37%	Asian/PI	11%
		Black	15%
		Hispanic	0%
		White	70%
		Other	4%

Training on the Modified Briefing Book Method

The standard setting began with Achieve providing the context of the ADP Algebra II EOC exam which included discussing the research findings that many students are not prepared for credit-bearing college courses and defining college-readiness for the purpose of the standard setting. In addition, the creation of the exam was mentioned including the purpose of the exam and its rigor. Next, Pearson provided an overview of the purpose of standard setting in general and a description of the process to be used in setting standards via the modified briefing book method. It was explained that the modified briefing book method was developed to consider the validity research evidence and emphasize the relationship between scores on the exam and performance in postsecondary education, rather than focus on item level data as is traditionally done with other standard setting methods. Hallmarks of the modified briefing book method included:

- Empirical data about the validity of the exam was provided in the form of a briefing book;
- Panelists explicitly considered the policy implications of the full set of data;
- Panelists reviewed the operational test to consider the context of the exam ;
- Panelists participated in a deliberative process with other stakeholders to provide basis for policy-based cut score recommendations; and
- Panelists took into account that multiple pieces of evidence and multiple lenses needed to be used to set a standard that is appropriate and useful given the multiple purposes of the exam.

The layout of the briefing book was reviewed and discussed as a large group. Following this general introduction to the method, panelist individually reviewed the operational assessment and discussed the rigor of the assessment with their table-mates.

Reviewing the Exam

An operational item book containing items from the most recent operational administration, spring 2009, was provided to panelists to review. The intent was to familiarize panelist with the content and rigor of the exam but not to have them thoroughly examine every item. A scoring key was provided that contained the calculator classification, item type, standard, points, possible, and the key of the item. This task was followed with a review of each of the validity studies included in the briefing book.

Reviewing the Briefing Book

The briefing book was a three-ring binder that contained various reports on validity research conducted in support of Algebra II EOC exam and the standard setting process. The purpose of the validity studies was to better understand how the ADP Algebra II Exam fits into the current landscape of mathematics instruction and assessment across public high schools and two- and four-year public colleges. Although the briefing book was mailed to all panelists in advance of the standard setting meeting, a considerable portion of the meeting was devoted to reviewing and discussing each validity study that was contained in the briefing book. Acknowledging that all studies included some limitations, the strengths and weaknesses of the study including the sample of students, the instruments used, and the methods implemented were discussed. The briefing book was separated into eight sections:

1. Briefing Book Overview: overview of the project and content of the briefing book
2. Concurrent Studies: executive summaries and full research reports for both the national and state-level studies.
3. Cross-Sectional Studies: executive summaries and full research reports for both the predictive and contrasting groups studies.
4. Judgment Studies: executive summaries and full research reports for all three regional meetings.

5. Mapping to Performance Level Descriptors: explanation of process and summary of results.
6. Cross-Walk of Cut Scores: explanation and summary spreadsheet of pertinent study outcomes.
7. Appendix A: explanation of additional technical details related to the studies and their summaries.
8. Appendix B: content alignment studies as written by Achieve.

After spending time reviewing and discussing each validity study in detail, panelists engaged in three rounds of recommending cuts for the “prepared” and “well prepared” performance levels.

Recommending Cut Scores

Prior to making their first round of recommendations, panelists were provided a survey which asked them to describe the features they believed were most important for discriminating among the three performance levels and which evidence they considered the most influential as well as what policy implications they were considering while making their recommendations. This ‘reflection’ survey was aimed at allowing them time to absorb and reflect on the various sources of information provided to them and the implications pertinent to the panelist’s role as either a state representative or a higher education representative. The goal was to ground their deliberations of the data and policy considerations in a more concrete fashion before making initial recommendations.

Also prior each round of recommendations being made, panelists filled out ‘readiness surveys’ stating that they understood the task at hand and were comfortable with proceeding with their task. After the second round, room level discussion occurred regarding the overall group results and then impact data was presented. Panelists were allowed a third and final round of ratings and then returned all materials, including a demographic survey, “evidence considered” surveys and an exit survey. The agenda for the standard setting event is shared in Appendix E.

Panelists provided their recommended cut scores using a raw score metric (0 to 76 possible points) based on their evaluation of the data provided in the validity studies along with their curriculum expertise and/or policy concerns. Feedback in terms of the mean, median, standard deviation, minimum, and maximum cut scores were presented after each round. This data was presented as disaggregated by role (state representative vs. higher education professional) and for combined groups (table-level or whole room, depending on round). Following the first round of ratings and presentation of feedback, the panelists engaged in table-level discussions about the resulting cuts for their table as well as their individual cut scores. Panelists were informed that a consensus on their judgments was not intended, but rather discussion for why differences exist was the goal.

Following this first round of recommendations, feedback and table-level discussion, panelists engaged in a second round of recommendations. Feedback following round 2 recommendations included the panelist’s individual cut scores, table-level cut scores and room-level cut scores. Panelists were allowed time to discuss these new

recommendations and then impact data were provided. The impact data were summarized by a graphic representation of what percentages of students would be at each performance level if the cut scores from the current round were applied to the 2009 population of test takers for various populations including: all students, ethnic subgroups, and gender subgroups. Panelists were reminded that the impact data was intended to inform but not dictate their ratings and that their final round 3 recommendations for the cut scores should be based on the goals of the exam, the specific information contained in this briefing book, discussions that occurred during the standard setting meeting, and their own best judgment.

Following the presentation of the impact data and resulting discussion, panelists filled out a final reflection survey prior to making their final round three recommendations. Upon completing round 3, panelists were asked to fill out an exit survey addressing their comfort and understanding of the procedures used along with a survey addressing what evidence they relied most heavily on in their deliberations as well as a demographic survey and reimbursement form.

Standard Setting Results

This next section presents the results of the three rounds for the overall group as well as broken down by type of panelist: higher education representatives and state representatives. Each table presented shows the mean, median, standard deviation, minimum and maximum cut scores assigned within each round. Recall that cut scores were made on the spring 2008 exam which had a raw score range of 0 to 76.

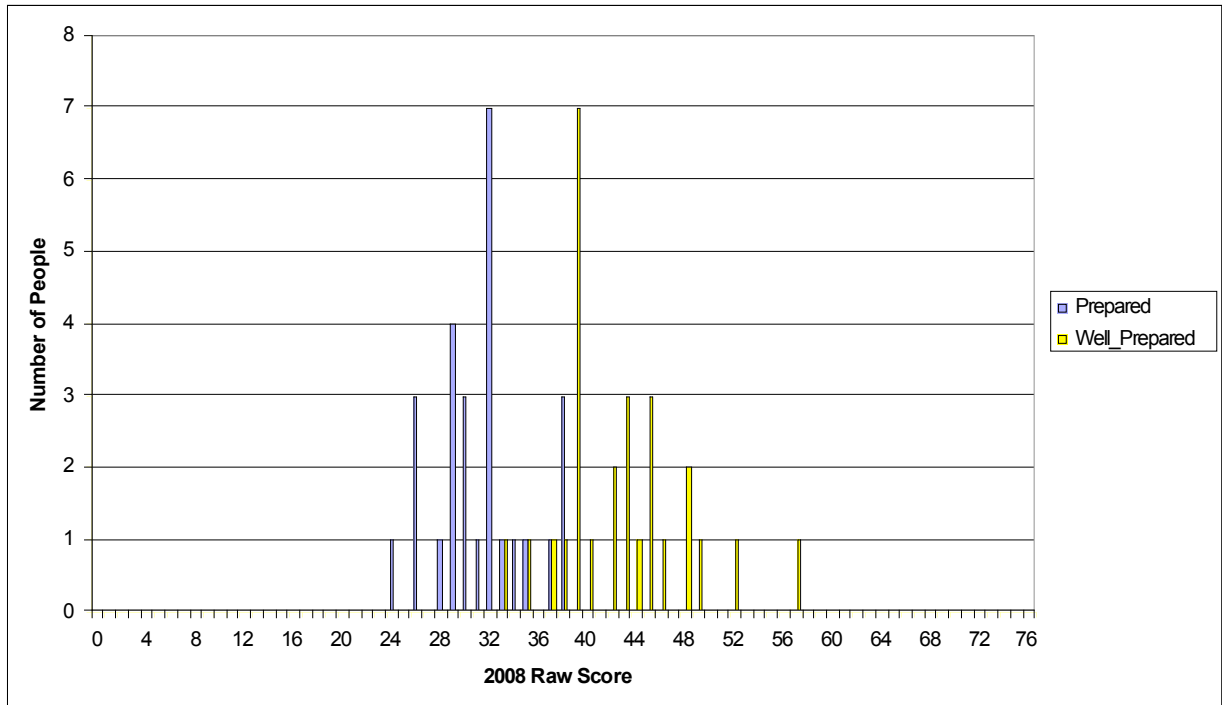
Round 1 Overall. The results for round 1 indicate fair distinction between the two cut scores. The means of the prepared and well prepared cuts were 31 and 43, with standard deviations of 3.8 and 5.3, respectively.

Table 2. Round 1 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	31.3	32.0	3.8	24	38
Well Prepared	42.5	42.0	5.3	33	57

Figure 1 displays the two distributions of cuts. There is some overlap between the two distributions which are both concentrated in the middle of the scale.

Figure 1. Round 1 Raw Score Cut Distributions



Round 1 by Panelist Type. Higher education representative and state representatives had similar cuts, although the well prepared cuts of the state representatives were slightly higher as shown in Tables 3 and 4.

Table 3. Higher Education Representatives' Round 1 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	31.1	32.0	4.5	24	38
Well Prepared	41.6	39.5	6.6	33	57

Table 4. State Representatives' Round 1 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	31.4	31.0	3.2	26	38
Well Prepared	43.3	43.0	4.1	39	52

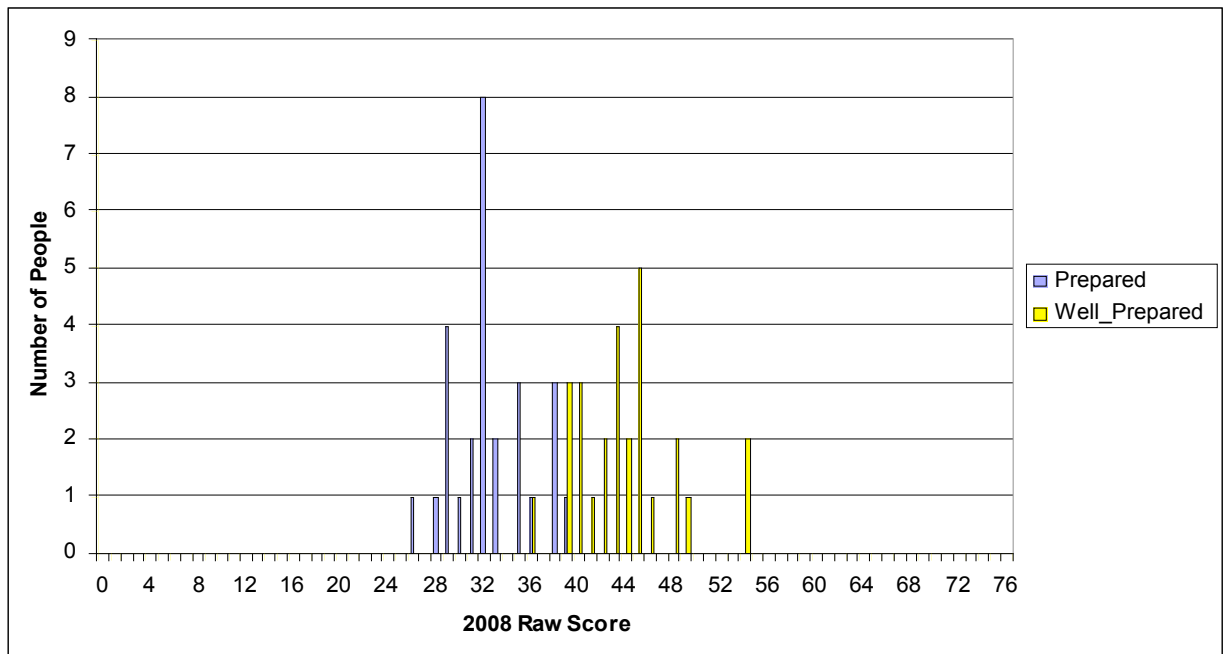
Round 2 Overall. Table 5 displays the results for round 2 which show little change from round 1. The median of the prepared cut remained at 32 which the well prepared cut increased slightly to 43. The standard deviation of the well prepared cut decreased, indicating a higher agreement among panelists.

Table 5. Round 2 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	32.5	32.0	3.3	26	39
Well Prepared	43.8	43.0	4.3	36	54

Figure 2 displays the narrowing of the greater distinction between the two distributions of the round 2 as compared to round 1.

Figure 2. Round 2 Raw Score Cut Distributions



Round 2 by Panelist Type. The mean and median raw score cuts of state representatives and higher education representatives are remarkably similar, as shown in Tables 6 and 7. However, there is slightly greater variation in the ratings of the higher education representatives.

Table 6. Higher Education Representatives' Round 2 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	32.5	32.5	4.0	26	39
Well Prepared	43.5	43.5	5.1	36	54

Table 7. State Representatives' Round 2 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	32.5	32.0	2.9	28	38
Well Prepared	44.0	43.0	3.6	39	54

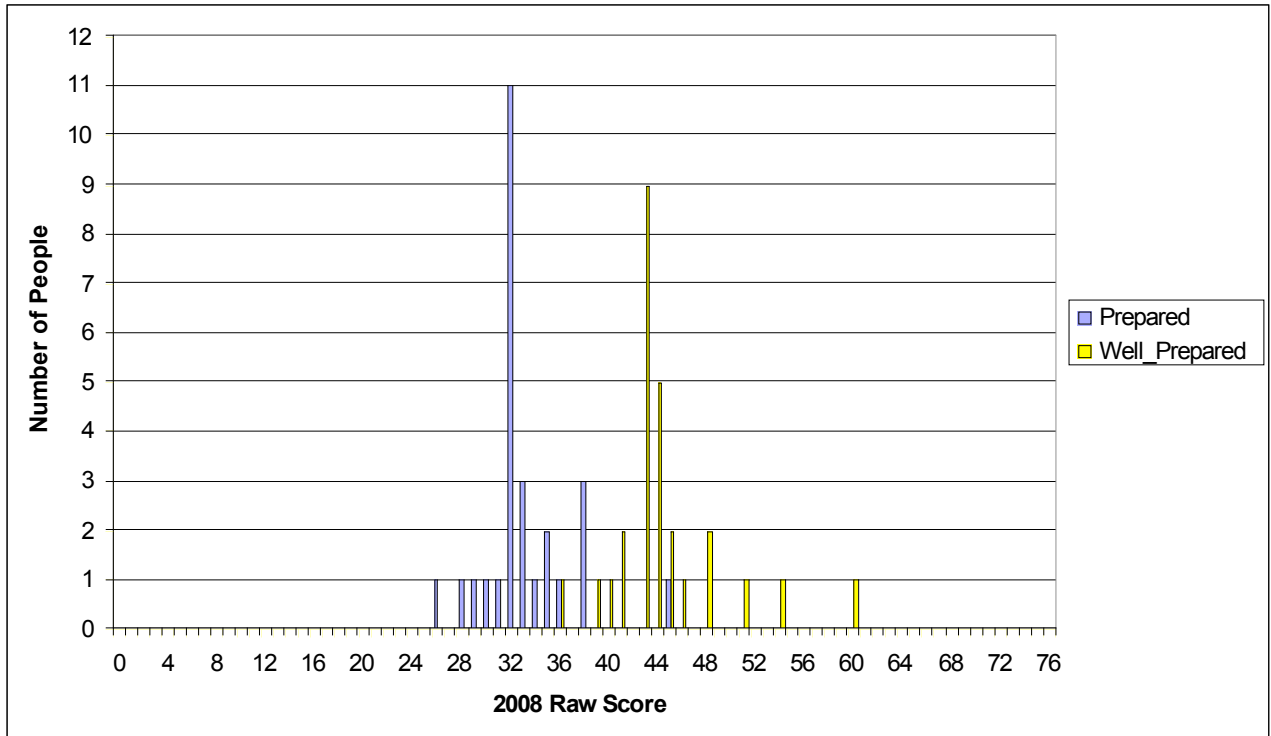
Round 3 Overall. Compared to round 2, there was a slight increase in means of the two cuts but the medians remained at 32 and 43, respectively. In addition, the maximums and standard deviations increased, suggesting less agreement among panelists.

Table 8. Round 3 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	33.1	32.0	3.7	26	45
Well Prepared	44.5	43.0	4.7	36	60

Figure 3 displays the two distributions of cuts for round 3. There appears to be one outlier at 60 which was not present in round 2.

Figure 3. Round 3 Raw Score Cut Distributions



Round 3 by Panelist Type. While the mean and median cuts of higher education representatives and state representatives were similar, the agreement among state representatives was stronger. This is evident in the smaller ranges and lower standard deviations of the state representatives' ratings, as shown in Tables 9 and 10.

Table 9. Higher Education Representatives' Round 3 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	33.6	32.5	4.7	26	45
Well Prepared	44.5	43.5	6.0	36	60

Table 10. State Representatives' Round 3 Results in Sp 2008 Metric

Performance Level	Mean	Median	Std Dev	Minimum	Maximum
Prepared	32.7	32.0	2.7	28	38
Well Prepared	44.5	43	3.6	39	54

Final Recommended Cut Scores. The final recommended cut scores resulting from the Modified Briefing Book procedure for Prepared and Well Prepared were 32 and

43 (of 76 points), respectively. It was acknowledged by the panelists that these cuts reflect the rigor of the ADP assessment frameworks and associated assessment.

Evidence Survey Results. Before rounds 1 and 3, panelists filled out a reflection survey in which they identified the evidence that would be most influential in their deliberation of recommending the cut scores. Results are shown below in Table 11. Keep in mind that most panelists based their decisions on several pieces of the evidence. In both rounds the judgment studies appeared to influence the majority of panelists. Many participants also chose the predictive studies (52%) and the concurrent studies (52%) in round 1 but these studies were of less influence in round 3.

Table 11. Evidence Survey

Evidence	Percent of Panelists	
	Round 1	Round 3
Judgment Studies	78%	78%
Predictive Studies	52%	44%
Concurrent Studies	52%	33%
PLD Mapping	22%	22%
Content	11%	11%
Other	7%	11%

Exit Survey Results. Following all activities, panelists were asked to indicate their level of agreement to each of the eight statements below. The overall results indicate that the majority of the panelists either ‘Agreed’ (4) or ‘Strongly Agreed’ (5) with each of the statements. Following each statement is the average agreement level on the 5-point Likert scale. For the complete breakout of results by level of agreement for the group, please see Appendix C.

1. I reviewed the Briefing Book contents in advance of the meeting ($\bar{x} = 4.7$).
2. I found the overview of the development and purpose of the ADP Algebra II End-of-Course Exam helpful in providing context to making my recommendation ($\bar{x} = 4.6$).
3. I found the information in the briefing book to be helpful in making my recommendation ($\bar{x} = 4.1$).
4. The discussion of the contents of the briefing book with the other panelists was helpful in making my recommendation ($\bar{x} = 4.5$).
5. I found the feedback on my cut score compared to other panelists useful in setting the standards ($\bar{x} = 4.2$).
6. I found the impact data and related discussion following round 2 judgments useful in setting the standards ($\bar{x} = 3.5$).

7. I am confident that my final cut score recommendation reflects the ability level of the “just barely **Prepared**” student ($\bar{x} = 4.2$).
8. I am confident that my final cut score recommendation reflects the ability level of the “just barely **Well-Prepared**” student ($\bar{x} = 4.1$).

Discussion

In general, the modified briefing book method was deemed a success. Panelists rated the standard setting process and resulting outcomes highly and were actively and often times enthusiastically engaged in all tasks and discussions throughout the standard setting meeting. Some limitations encountered along the way related to the generalizability of the data we collected. Given that not all states participating in the consortium were involved in each study, nor were any data collected randomly (convenience only), it was important to note the limitations of the data in each research report and to reemphasize this at each round of the standard setting process to adequately forewarn panelists from relying too heavily on any single study. Knowing these limitations existed, discussions of additional future research are already underway to evaluate the appropriateness of the standards (perhaps, in part, by encouraging and assisting the ADP Network consortium states with following students through to college).

This approach to setting standards seems poised to be used more frequently in the future with increasing emphasis being placed on college and workplace readiness and the implementation of common core assessments. These settings will call for an increased focus on interpreting assessment results in relationship to external variables and create situations in which cut scores must be based on supporting validity research in addition to, or perhaps in place of, the traditional content focus. For example, the National Assessment Governing Board is planning research and validity studies that will enable the National Assessment of Educational Progress (NAEP) to report on the preparedness of 12th graders for postsecondary education and job training after they graduate from high school (National Assessment Governing Board, 2008). The Modified Briefing Book approach developed and executed in this study will provide the assessment community with a concrete example of how validity data can be organized and structured to facilitate a fully informed and participatory standard setting process.

References

- Achieve, Inc. (2008). American diploma project algebra II end-of-course exam: 2008 annual report. Retrieved July 30, 2009 from <http://www.achieve.org/files/ADPAlgebraIIEnd-Of-CourseExam2008AnnualReport.pdf>.
- Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. Educational Measurement: Issues and Practice, 21 (1), 16–22.
- Haertel, E. H. (2008). Standard setting. In K. E. Ryan and L. A. Shepard (Eds.), The future of test-based educational accountability (pp. 139-154). Routledge.

National Assessment Governing Board (2008, November). Technical panel on 12th grade preparedness research. Final report, pre-publication edition. Retrieved July 31, 2009 from <http://www.nagb.org/newsroom/PressReleasePDFs/12grade-preparedness-report.pdf>.

Appendix A

Executive Summary - Concurrent Validity Studies

Background

- The concurrent validity studies provide data about how the ADP Algebra II End-of-Course Exam relates to other mathematics assessments.
- Two types of exams were included in the concurrent validity studies:
 - National exams, specifically the math sections of the ACT, SAT, and PSAT assessments.
 - State exams, specifically mathematics assessments administered at the high school level in six different states.
- The concurrent validity studies provide indirect information for considering the use of the ADP Algebra II exam to assess college readiness. For example, the SAT and ACT already provide implicit and explicit benchmarks of college readiness. The PSAT is used as a qualifying exam for the National Merit Scholarship program, and the state tests classify students into “Proficient” and “Advanced” performance categories.
- The relevance of the various concurrent validity studies for the ADP Algebra II standard setting varies for a number of reasons. First, the exams studied vary in how closely they match the content measured by the Algebra II exam. Second, the exams differ in how well performance on them is thought to relate to college readiness. Finally, the amount and quality of the data collected across the various studies differed.

Data

- National exam scores matched with spring 2008 ADP Algebra II exam scores:
 - ACT scores were matched in Arkansas for 6,278 students and in Kentucky for 999 students.
 - SAT scores were matched in Indiana for 205 students and in Pennsylvania for 414 students.
 - PSAT scores were matched in Rhode Island for 954 students.
- State exam scores matched with spring 2008 ADP Algebra II exam scores:
 - Hawaii State Assessment (HSA) scores were matched for 1,164 students.
 - Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) scores were matched for 2,858 students.
 - Kentucky Core Content Test (KCCT) scores were matched for 923 students.

- New Jersey High School Proficiency Assessment (HSPA) scores were matched for 206 students.
- Pennsylvania System of School Assessment (PSSA) scores were matched for 3,015 students.
- New England Common Assessment Program (NECAP) scores were matched for 882 students in Rhode Island.

Methods

- Spring 2008 ADP Algebra II exam scores and scores on the national and state exams were linked using equipercentile linking, resulting in a concordance table for each exam.
- For the ACT and SAT, linear regression was conducted to predict ACT/SAT scores exam scores from spring 2008 ADP exam scores.
- For the ACT and SAT, logistic regression was conducted to investigate the probability of earning a particular ACT/SAT score, given an ADP exam score.

Results

- ACT
 - Strong linear relationship between ADP exam and ACT ($r = .698$).
 - The ACT “college-readiness” score of 22 is associated with spring 2008 ADP exam scores of 25 and 26.
 - An ADP exam core of 32 is associated with having a 65% chance of earning a 22 or better on the ACT.
 - An ADP exam score of 31 is associated with a predicted ACT score of 22.
- SAT
 - Moderate correlation between ADP exam and SAT ($r = .490$).
 - Using the concordance table between ACT and SAT, the ACT “college-readiness” score of 22 is associated with SAT scores of 520 and 530. These scores mapped to ADP exam scores of 23 and 24.
 - ADP exam scores of 36 and 39 are associated with a 65% chance of earning SAT scores of 520+ and 530+, respectively.
 - An ADP exam scores between 36 and 40 linearly predict SAT scores in the 520-530 range.
- PSAT
 - Strong linear relationship between ADP exam and PSAT ($r = .616$).
 - The level of PSAT math performance associated with being recognized as a National Merit Scholar corresponds to ADP exam scores of 56 to 59.
- State exams
 - Mean scores on the spring 2008 ADP exam for the matched students varied considerably by state, ranging from 13.8 to 25.1.

- Correlations between ADP exam and the state exams also varied, ranging from .323 (New Jersey HSPA) to 0.712 (Indiana STEP+).
- Based on the concordance tables, the “proficient” levels on the state exams are associated with ADP exam scores ranging from 12 to 20.
- Based on the concordance tables, the “advanced” levels on the state exams are associated with ADP exam scores ranging from 21 to 43.

Considerations

- The content match between the exams included in the concurrent studies and the Algebra II exam is limited.
 - The ACT, SAT, and PSAT contain varying amounts of algebra content, little of which is likely to be at the Algebra II level.
 - The state exams are primarily used for NCLB and/or as part of graduation requirements, and are unlikely to measure Algebra II content.
- The quality of the concurrent data varies across the studies.
 - The ACT sample is fairly representative of the ACT testing population.
 - The SAT sample is lower performing than the SAT testing population.
 - The PSAT sample is representative of the PSAT testing population in terms of mean performance, but is significantly less variable.
 - The state exam samples vary in their quality and the ability levels of the students matched to the ADP Algebra II exam.
- The ACT concurrent validity study seems most relevant for the purposes of the ADP Algebra II standard setting.
 - The ACT sample is largest and ACT scores are highly correlated with ADP Algebra II scores.
 - ACT’s definition of college-ready is having a 50% chance of earning a B or better and a 75% chance of earning a C or better in College Algebra. ACT’s benchmark for this is an ACT mathematics score of 22 or higher.
- The SAT concurrent validity study is relevant for the purposes of the ADP Algebra II standard setting; however, the SAT results should be interpreted with caution.
 - There is no established “college-readiness” score for the SAT.
 - The SAT sample was small and unrepresentative of the SAT testing population.
 - Correlations with ADP Algebra II scores in the concurrent validity study were lower than would be expected.
- The PSAT study provides validity data for the ADP Algebra II exam, but is less relevant for the standard setting.
 - Although a high correlation between PSAT and ADP scores was found, there is not a clear relationship between PSAT and “college-readiness.

- The concordance for the benchmark score associated with eligibility for a National Merit Scholarship provides some indication of an upper level of performance on the ADP Algebra II exam; however, this level of performance likely exceeds the level of performance associated with college readiness.
- PSAT data were limited to students in Rhode Island.
- Although they provide validity evidence for the ADP Algebra II exam, the studies involving state exams are probably least relevant to setting the college readiness standards.
 - The tests largely measure different mathematics skills at a different level of rigor.
 - Connections between proficiency levels on the state tests and inferences of college readiness are unknown.

Appendix B

Executive Summary – Cross-Sectional Validity Studies

Background

- The cross-sectional validity studies provide data about how college students perform on the ADP Algebra II exam. Students in Algebra and Pre-Calculus courses from community colleges, 4-year typical institutions, and 4-year more selective institutions participated.
- The data collected for the cross-sectional validity studies were analyzed in two different ways:
 - Predictive Study-examined how students' ADP exam score predicts their course grade and provides information regarding what ADP scores seem to signify college readiness based on the final course grades of students. For this analysis, only Algebra and Pre-Calculus students with assigned final grades were included.
 - Contrasting Groups Study-examined how the distributions of scores for students earning a B or better differ from those students who earn less than a B. Resulting cut scores provide information regarding the scores that best separate these distributions.

Data

- The exam was administered at the beginning of the semester and scores were matched with final course grades.
- Data were collected in Fall 2008 and Spring 2009 from 31 institutions.
- 3,132 students had matching ADP exam scores and final course grades.
- Students were split fairly evenly between institution types, but the number of students in Pre-Calculus was less than half the number of those in Algebra.

Methods

- Predictive-For each institution type by subject subgroup, logistic regression was conducted to investigate the score needed in order to have a 65% chance of earning an A or better, B or better, and C or better in the student's first credit-bearing course.
- Contrasting Groups-Within each institution type, smoothed distributions for "successful" students (those earning a B or better) were compared to those students earning less than a B, and the intersections were determined.

Results

- Predictive
 - ADP Exam scores associated with a 65% probability of earning a B or better in an Algebra course varied by institution type, ranging from 32 to 48.
 - For Pre-Calculus, ADP Exam scores of 32 and 38 for community college and 4-year typical institutions, respectively, were associated with a 65% probability of earning a B or better. The model did not fit for 4-year more selective institutions.
- Contrasting Groups
 - For community college students, a score of 25 separated successful students from those earning less than a B.
 - For 4 year typical institution students, a score of 26 optimally successful students from those earning less than a B.
 - For 4 year more selective institution students, a score of 23 separated successful students from those earning less than a B.

Considerations

- Lack of motivation likely influenced students' scores.
- The restriction in range of ADP Algebra II scores affected the predictive study results.
 - For the community college Algebra sample, only three out of 761 students received scores of 48 or higher.
 - For Algebra students from 4-year more selective institutions, the highest score was a 42.
 - Relatively small differences in mean Algebra II scores were seen for students receiving different grades, especially in the data based on Algebra courses.
- The contrasting groups' analyses are limited by the following considerations.
 - The sample size for each distribution is relatively small sample sizes (~500 students).
 - Restriction of range of ADP Algebra II scores also affected the contrasting groups' results.
- The logistic regression results reported in the predictive studies seemed more credible for the Pre-Calculus groups than the Algebra groups.
 - For the Pre-Calculus groups, ADP Algebra scores between 32 and 38 are associated with a 65% probability of earning a B or better.
 - The logistic regression results for the community college Algebra sample should probably be discounted.
- The 65% probability value used with the logistic regressions significantly impacted results to other probability values that might be considered. For

example, a 50% probability of earning a B or better seemed to be associated with ADP Algebra scores between 20 and 28.

Appendix C

Executive Summary – Judgment Study

Background

- The judgment study provides data about how the ADP Algebra II exam relates to college professors' expectations of what students need to have previously mastered in order to successfully learn the material that will be covered in their College Algebra or Pre-Calculus course, where "successful" is defined as earning a B or better without remediation.
- The data gathered were used to inform the ADP Algebra II content-based performance level descriptors for "Needs Intensive Preparation", "Needs Preparation", "Prepared", and "Well Prepared" and to calculate an initial cut score separating "Prepared" students from "Needs Preparation" students.
- At the time of the judgment studies, four performance levels were planned; however, the lowest level "Needs Intensive Preparation" has since been dropped.

Participants

- Three one-day meetings were held in the spring of 2009.
 - Little Rock, Arkansas
 - Baltimore, Maryland
 - Cleveland, Ohio
- 133 professors representing 79 institutions from 20 states participated.
- Professors were currently teaching either Algebra or Pre-Calculus
- Community college, 4-year typical institutions, and 4-year more selective institutions were represented

Methods

- Participants were presented with three major tasks:
 - Standards Relevance Survey-Each ADP Algebra II benchmark was rated as to its relevance in preparing successful students for the first college level credit-bearing math course. Benchmarks were rated as Essential, Important, Helpful, or Not Relevant.
 - Defining Threshold Descriptors-Participants described what knowledge, skills, and abilities students in each of the performance levels would be able to demonstrate on the first day of their college-level math course.
 - Item-Level Judgments-Participants examined each item in the Spring 2008 operational exam and determined the number of points a "successful" student would earn on the first day of class.

Results

- Standards Relevance Survey

- There was clear agreement that Real Numbers and Algebraic Expressions represent important or essential skills that should be previously mastered by incoming students in order to be successful in their first college level math course.
- Elements of Linear Equations and Inequalities, Quadratic Functions, and Function Operations were also considered important skills by most participants.
- Defining Threshold Descriptors
 - See Tab 5 for the final Performance Level Descriptors
- Item-Level Judgments
 - Following Round 2, cut scores varied by course and institution type with subgroup medians ranging from 23.0 to 38.0.
 - Pre-Calculus cut scores were consistently higher than Algebra cut scores.
 - Cuts for the 4-year institutions were higher than those for community colleges.
 - The median cut score from all data was 29.0.

Considerations

- While the instructors all taught College Algebra or Pre-Calculus the content varied considerable across the institutions. (See Appendix B3: Analysis of Judgment Study Syllabi.)
- The understanding of a “B or better” student was likely interpreted differently by each instructor based on their own individual expectations and the grading norms at their institutions.
- Anecdotal evidence from comments made by the instructors at the meetings suggests that the instructors have lowered their expectations to accommodate an increasingly unprepared population of students (i.e., “I don’t expect them to know this since I have to teach it anyway”).

Appendix D

Cross-Walk of Cut Scores

The Crosswalk document was created to summarize and illustrate the results of the various validity studies. The first column contains the Spring 2008 raw score scale which ranges from 0 to 76. The next six columns present results of the respective validity studies. The cells shaded in orange represent results that are more relevant for standard setting purposes, as recommended by the technical advisors to the ADP assessment consortium. A key for the abbreviations used is shown below:

AL = Algebra

PC = Pre-Calculus

CC = Community College

4T = 4-year Typical Institution

4S = 4-year More Selective Institution

Spring 2008 Raw Score	State Concurrent (Proficiency Levels)	National Concurrent	Predictive Study	Contrasting Groups (Predictive data)	Judgment Studies	Mapping to PLDs
0						
1						
3						
4						
6			C or better in CC PC			
7						
8						
9						
11						
12	NJ-Proficient		C or better in CC AL			
13	IN-Proficient					
14	HI-Proficient					
16	KY & PA-Proficient		C or better in 4S AL			
17						
18						
19						
20	RI-Proficient					
21	HI-Advanced		C or better in 4T PC			
23	NJ & PA-Advanced	SAT-Concordance		4-Year Selective	CC AL Cut	
24	KY-Advanced	SAT-Concordance			All CC Cut & All AL Cut	
25		ACT-Concordance		Community College		
26		ACT-Concordance		4-Year Typical	4S AL Cut	
27					All AL Cut	
29	IN-Advanced				4T AL Cut & All PC Cut	
30						
31		ACT-Predicted Score			All 4S Cut	
32		ACT-Expectancy Table	B or better in 4S AL & CC PC			
33					CC PC Cut & All 4T Cut	Prepared

Spring 2008 Raw Score	State Concurrent (Proficiency Levels)	National Concurrent	Predictive Study	Contrasting Groups (Predictive data)	Judgment Studies	Mapping to PLDs
35					All PC Cut	
36		SAT-Exp. & Pred. Score				
37		SAT-Pred. Score	B or better in 4T AL			
38		SAT-Pred. Score	B or better in 4T PC		4T & 4S PC Cut	
39		SAT-Exp. & Pred. Score				
41	RI-Advanced	SAT-Exp. & Pred. Score	A or better in 4S AL			
42	RI-Advanced	SAT-Exp. & Pred. Score				
43	RI-Advanced					
44						
45						Well Prepared
46						
48			B or better in CC AL			
49			A or better in 4T AL			
50			A or better in CC PC			
51						
52			A or better in 4T PC			
53						
54						
55						
56		PSAT-Concordance				
57		PSAT-Concordance				
58		PSAT-Concordance				
59		PSAT-Concordance				
60						
:						
76						

Appendix E

Algebra II Standard Setting Agenda

Day 1	(Wednesday, July 22, 2009)
8:30-9:00	Registration/Breakfast
9:00-9:30	Opening Remarks & Introductions
9:30-10:45	Overview <ul style="list-style-type: none">• The American Diploma Project (<i>Achieve</i>)• The ADP Algebra II End-of-Course Exam (<i>Achieve</i>)• ADP Algebra II Exam Standard Setting Approach• Validity Studies to Inform Standard Setting
10:45-11:00	Break
11:00-12:00	Review the Spring 2009 Operational Test Items
12:00-1:00	Lunch
1:00-2:00	Review and Discuss Results from the Concurrent Validity Studies <ul style="list-style-type: none">• National Exams• State Exams
2:00-3:00	Review and Discuss Results from Cross-Sectional Studies <ul style="list-style-type: none">• Predictive Studies• Contrasting Groups
3:00-3:15	Break
3:15-4:30	Review and Discuss the Results of the Judgment Studies
Day 2	(Thursday, July 23, 2009)
8:30-9:00	Breakfast
9:00-9:30	Mapping to Performance Level Descriptors
9:30-10:00	Cross-Walk of Cut Scores
10:00-10:30	Recap of All Data
10:30-10:45	Break
10:45-12:00	Make Round 1 Judgment on Cut Scores
12:00-1:00	Lunch
1:00-1:30	Small group discussion of table agreement data from round 1
1:30-2:00	Make Round 2 Judgments
2:00-2:30	Break
2:30-3:00	Large group discussion of agreement from round 2
3:00-4:00	Present and discuss impact data from round 2
Day 3	(Friday, July 24, 2009)
8:30-9:00	Breakfast
9:00-10:00	Make Round 3 Judgments
10:00-10:30	Large group discussion of agreement and impact data from round 3
10:30-11:00	Complete exit survey
11:00-11:30	Check-in materials, adjourn for CDT meeting
11:30-12:30	Break/Lunch