Design and Evaluation of Mixed-Format Large Scale Assessments

Rosemary Reshetar & Gerald J. Melican
The College Board

Abstract

This paper discusses issues related to the design and psychometric work for mixed-format tests—tests containing both multiple-choice (MC) and constructed-response (CR) items. The issues of validity, fairness, reliability and score consistency can be addressed but for mixed-format tests there are many decisions to be made and no examination or examination program faces exactly the same choices. This paper raised some issues and used the Advanced Placement Program® (AP®) experience to illustrate the types of questions that need to be addressed and outlined the strong test development processes that need to be in place to address the issues. The diversity of the 34 AP exam designs provides a unique opportunity to review and investigate the test design and psychometric issues that accompany mixed format testing in a high stakes setting.

The AP Program

33 Courses and exams in 22 subject areas

AP is a full package including curriculum guidelines, professional development opportunities for teachers, and examinations

Experienced continuous growth since inception in 1955

- •2009: 2.9 million exams
- •1999: close to 700,000 exams

Most colleges and universities in US, along with institutions in more than 40 other countries, use AP Exam results to award credit or allow placement into a higher-level course

See College Board's AP Central Website for information.

The AP Exams

Each AP course has a corresponding exam that participating high schools worldwide administer in May in a secure environment.

Except for the three Studio Art Exams, which are portfolio assessments, AP Exams contain both multiple-choice (MC) and constructed-response (CR) sections.

The **contribution of the constructed-response section** performance varies across the examinations ranging from **33 to 60 percent** in terms of the number of points on the composite raw score.

Constructed-response items differ greatly over examinations and include

- •essays,
- •short answers,
- •document based questions,
- •graph creation,
- •problem based questions,
- •speaking
- •aural/nonaural, and sight singing items.

Scoring for some of these item types is holistic, for other item types, analytic, and for some a combination of analytic and holistic.

Features and 2009 Volumes of AP Exams

Subject	2009 Volume	Special Features	Exam Design CR
Arts			
Art History	20,619	Until 2009 Slides (Some MC & CR); 2010 printed inserted	Essay; short essays with picture prompts
Music	15,438	Listening & Sight Singing (Recorded)	Music writing: dictation with listening prompts; sight singing
Studio Art- Drawing	14,589	Portfolio	3 sections: Quality (artwork submitted); Concentration and Breadth (digital submission)
Studio Art -2D Design	17,387	Portfolio	3 sections: Quality (artwork submitted); Concentration and Breadth (digital submission)
Studio Art - 3D Design	2,761	Portfolio	3 sections: Quality, Concentration and Breadth (digital submission)
English			
English Literature	332,352	Internal choice, unlimited	3 Essays
English Language	337,441	1 Synthesis essay	3 Essays
World Languages			
Chinese Language & Culture	5,100	Listening & Speaking; Computer Delivery	Presentational & interpersonal speaking and writing
French Language	21,029	Listening & Speaking	Writing and speaking
German Language	5,001	Listening & Speaking	Writing and speaking
Japanese Language & Culture	2,085	Listening & Speaking; Computer Delivery	Presentational & interpersonal speaking and writing
Latin Vergil	4,295	Internal choice (limited)	2 translations, 1 long essay, 2 short essays
Spanish Language	110,723	Listening & Speaking	Presentational & interpersonal speaking and writing
Spanish Literature	16,633	Internal choice (limited to 2); Content & Lang scores on essay	3 Essays
History			
European History	101,359	External choice essays	DBQ + 2 Essays
US History	360,173	External choice essays	DBQ + 2 Essays
World History	143,426	Internal choice essays (Limited to 2 or 3)	DBQ + 2 Essays
Math			
Calculus AB	230,588	Calculator (Some MC & CR)	Problem sets
Calculus BC	72,965	Calculator (Some MC & CR)	Problem sets
Computer Science A	16,622		Program coding
Statistics	116,876	Calculator	Short answer sets; 1 investigative task
Science			
Biology	159,580		Set-based short answer
Chemistry	104,789	Calculator (Some CR)	Problem-based sets; essay/short answer
Environmental Science	73,575		Set-based short answer
Physics B	62,702	Calculator for CR	Problem-based sets
Physics C: Mechanics	29,167	Calculator for CR	Problem-based sets
Physics C: Elec & Mag	12,628	Calculator for CR	Problem-based sets
Social Science			
Microeconomics	46,272		Sets with graph creation, short answer and short essay
Macroeconomics	73,817		Sets with graph creation, short answer and short essay
Comp Gov & Politics	14,728		Short essay/short answer sets
US Gov & Politics	189,998		Short essay/short answer sets
Human Geography	50,730		3 Short answer/short essay sets
Psychology	151,006		2 Essays

Fairness

Aspects of interest for this paper are differential performance across item types, coachability/test preparation, and security.

Mantel-Haenszel DIF for MC items
CR and MC items undergo fairness review

Coachability and test prep
•Coachability may be less of a concern for AP, because of the heavy emphasis on content and skills
•CR items disclosed after the administration and posted on CB website along with rubrics and sample responses

Security
•CR items could be memorable, but disclosed. However this means they can't be used for NEAT equating.

Validity

Relevant validity research focused on student outcomes such as academic success in college improved graduation rates.

Ewing (2006) summarized studies. that have investigated the impact of the AP program on student outcomes such as grade point average in subsequent courses, retention, and graduation. For example, Dodd, Fitzpatrick, De Ayala, and Jennings (2002) reported that students who were exempt from courses due to their AP score performed as well or better in subsequent courses and took more courses in the relevant content area. Dougherty, Mellor, and Jian (2006) reported that students earning a 3 or better on at least one AP exam were more likely to graduate from college in five years or less compared to non-AP students after controlling for prior academic achievement and school-level variables.

Reading Logistics

HUGE process which is integral to the AP Program; all CR item responses scored **on-site**
•Approximately 2 weeks each June
•About 1 week per exam

>10,000 readers per year (35 to 1,100 per exam)

Significant networking and professional development for readers; garners understanding of AP program and continued support of AP

AP Scores 1- 5

AP Exam grades are reported on a **1–5 scale**. Students' raw scores on the MC and CR items are weighted and combined. The weighted raw composite score is then converted to a grade on the AP 5 point scale.

AP Exam grades of 5 are considered equivalent to A grades in the corresponding college course; grades of 4 are equivalent to grades of A-, B+, and B and grades of 3 are equivalent to grades of B-, C+, and C in college.

Reliability

Exam scores of 1 through 5 should be consistent across forms
- Estimates of classification accuracy around .90.

Item and section level weighting schemes affect reliability; Balance optimizing reliability with SMEs evaluation of value of CR section. (See Swaminathan & Rogers and Hendrickson, Patterson & Ewing posters)

AP items scored by a single reader. Systematic training of readers as well as ongoing "backreading" to monitor performance is conducted. Given the reading logistics it would not be feasible to double read.

Special data collection used to routinely conduct reader reliability studies.

Stakeholder's View

AP credit and placement is primary value driver.

Post-secondary institutions must believe AP is a valid measure. The decision to use AP is at institutional level. Based on:

1) How the test looks.

Focus on CR items: Many SMEs believe that they are better able to test depth of knowledge and skill than MC. CR items and rubrics are published each year on AP Central website and are most noticeable.

2) How they believe the AP grades reflect performance in the corresponding course at their institution.

Influence

Exam Design

Structure (see Hendrickson, Patterson & Ewing poster)

Exam level aspects:
- MC and FR sections, weights
- Depth vs. breadth of content & skill coverage must be addressed since teachers' flexibility in varying the depth to which they explore the various topics is encouraged.

Item level aspects:
- Use of task models
- Types and consistency of scoring rubrics

Psychometricians prefer to emphasize score reliability, which could be increased by higher MC section weights. Yet higher CR section weights are valued by stakeholders.

Equating and Linking Plan

Considerations for equating and linking plan:

1. What are the form use needs?
   - Form level or item banking model
2. What data will be available to use for equating and linking?
   - Volume and appropriateness of equating sample
   - Timeline for production
3. What options are there for reuse of MC and CR items?
4. If cutscores are required, are there any additional steps needed to assure consistency of cutscores across forms?
   - Is equating sufficient?
   - Policy decision process needed

Take Away Message

Psychometricians don't rule the day.
At best we have veto power.
The user needs are top priority – as they should be.

Design and Evaluation of Mixed-Format Large Scale Assessments for the Advanced Placement Program® (AP®)

Rosemary Reshetar and Gerald J. Melican

The College Board

Paper presented at the Annual Meeting of the American Education Research Association in Denver, April 2010.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board presentations do not necessarily represent official College Board position or policy.

Abstract

This paper discusses issues related to the design and psychometric work for mixed-format tests—tests containing both multiple-choice (MC) and constructed-response (CR) items. The issues of validity, fairness, reliability and score consistency can be addressed but for mixed-format tests there are many decisions to be made and no examination or examination program faces exactly the same choices. This paper raised some issues and used the AP experience to illustrate the types of questions that need to be addressed and outlined the strong test development processes that need to be in place to address the issues.

Introduction

Constructed response (CR) items have become increasingly popular in large scale assessment programs, and many tests are composed exclusively of CR items or of a combination of CR and multiple-choice (MC) items; i.e., mixed format. Within the class of written response CR items, there exists a wide variety of task designs and scoring rubrics. At one end of the continuum for example are CR items that are of the fill-in nature and result in single best answers that are scored right or wrong. Other examples are open ended CR items that result in essays, results of experiments, and portfolios which are scored according to a holistic rubric. Bennett, Ward, Rock, and LaHart (1990) classified the variation among CR items according to the degree of constraint placed on the examinee's response.

Item types have strengths and weaknesses. MC items tend to provide better content coverage, and are efficient, reliable, cost-effective, and expedient to score. CR items on the other hand allow for measurement of skills that more closely resemble those valued in the curricular domain, match assessments in the college classrooms or behavior in a profession, and cover a wider range of content and cognitive demand. Recently, procedures and infrastructure capabilities, such as electronic scoring and distributed scoring of CR items, have allowed programs to more easily implement mixed format assessments and reap the benefits of both MC and CR items.

The test development and psychometric procedures for MC tests are well honed and best practices are commonly known. Some guidelines for approaching the constructed-response and mixed-format test development and psychometric procedures are available and quite helpful, see e.g., Welch (2006), Lane and Stone (2006), and Hogan and Murphy (2007). However, in the design of mixed-formats assessments and corresponding linking and equating plans there are many decisions that need to be made and there aren't obvious best answers to each. There are trade-offs all along the way, and the goal is to create the best measurement within the practical constraints of the program. This paper provides an overview of the issues related to the design, perpetuation, and evaluation of mixed-format examinations for high stakes examinations using the AP program as a real world example to highlight and elaborate mixed-format testing issues.

An overview of the AP program is presented next, followed by discussions of validity, fairness, reliability, and score consistency. Decisions concerning exam design are then discussed.

Overview of The Advanced Placement Program® (AP®)

Since its inception in 1955, the Advanced Placement Program® (AP®) has been a unique collaborative effort among motivated students, dedicated teachers, and committed high schools, colleges, and universities. AP is a full package including curriculum guidelines, professional development opportunities for teachers, and the examinations. The College Board partners with colleges and universities to create assessments of college-level learning—the AP Examinations.

The AP Program has experienced continuous growth and in 2009, 2.9 million exams were taken by nearly 1.7 million students at more than 17,000 high schools. Ten years ago these figures were close to 700,000 students at nearly 13,000 high schools. AP summary reports with participation rates can be found at http://professionals.collegeboard.com/data-reports-research/ap/data.

There are currently 33 courses and assessments in 22 subject areas. The College Board supports secondary schools in training teachers and developing a high quality curriculum of high academic intensity that enables students to meet the standards for college-level learning in these subjects. As a result, most colleges and universities in the United States, as well as institutions in more than 40 other countries, use AP Exam results to award credit or allow placement into a higher-level college course so that college entrants can move directly into the courses that match their level of academic preparation for college.

The AP Exams

Each AP course has a corresponding exam that participating high schools worldwide administer in May. Except for Studio Art, which is a portfolio assessment, AP Exams contain both multiple-choice (MC) and constructed-response (CR) sections. The CR sections vary across examinations include essays, problem solving, oral responses, and creating graphs. The contribution of the constructed-response section performance varies across the examinations ranging from 33 to 60 percent in terms of the number of points on the composite raw score. AP Exams represent the culmination of AP courses and are thus an integral part of the Program although not every AP student takes the examination and it is possible to take the examination without taking an AP course.

AP Exams differ in a variety of characteristics including content classification, year exam began (initiation date), recent national exam volumes, number of multiple-choice points possible and constructed-response points possible, percentage of composite based on multiple-choice items and constructed-response items, type of constructed-response scoring, the correlation between multiple-choice and constructed-response items, choice of constructed-response items, and length of exam. Table 1 provides a list of the exams and summary information about each.

Constructed-response items differ greatly over examinations and include essays, short answers, document based questions, graph creation, problem based questions, speaking, aural/nonaural, and sight singing items. Scoring for some of these item types is holistic, for other item types, analytic, and for some a combination of analytic and holistic. In a few cases, e.g., Spanish and French Literature, some constructed-response items are scored twice—once holistically for content and again holistically for language.

AP Exam grades are reported on a 1–5 scale. Students' raw scores on the MC and CR items are weighted and combined. The weighted raw composite score is then converted to a grade on the AP 5 point scale. The AP grade qualification definitions are:

5 Extremely well qualified;

4 Well qualified;

3 Qualified;

2 Possibly qualified; and

1 No recommendation.

AP Exam grades of 5 are considered equivalent to A grades in the corresponding college course. AP Exam grades of 4 are equivalent to grades of A-, B+, and B in college. AP Exam grades of 3 are equivalent to grades of B-, C+, and C in college.

Each AP exam is administered once during a two week period in May with the vast majority of test takers being administered the same form. Following that, each exam is offered once during a three-day late testing period with a different form. The scoring of the CR items is completed during the first three weeks of June and scores for all exams are reported by July 1. Within 48 hours after administration, the CR items are released and posted on the AP Central website.

The diversity of the AP exam designs provides a unique opportunity to review and investigate the test design and psychometric issues that accompany mixed format testing in a high stakes setting.

Validity

AP provides willing and academically prepared high school students with the opportunity to study and learn at the college level. More than 90 percent of four-year colleges and universities in the United States grant students credit, placement, or both on the basis of successful AP Exam Scores. As a result, research focused on student outcomes such as academic success in college and improved graduation rates is relevant to evaluation of AP's validity. Ewing (2006) summarized studies that have investigated the impact of the AP program on student outcomes such as grade point average in subsequent courses, retention, and graduation. For example, Dodd, Fitzpatrick, De Ayala, and Jennings (2002) reported that students who were exempt from courses due to their AP score performed as well or better in subsequent courses and took more courses in the relevant content area. Dougherty, Mellor, and Jian (2006) reported that students earning a 3 or better on at least one AP exam were more likely to graduate from college in five years or less compared to non-AP students after controlling for prior academic achievement and school-level variables.

AP's primary value driver for high school students and their parents is the acceptance and use of AP for credit, placement or admissions by post-secondary institutions. For AP to be of the greatest value, an institution will agree that performing at a certain level on an AP exam is an adequate substitute for taking the same or similar course at their institution. Also of high value is the use of AP exam scores for placement decisions. For each subject, this means that the post-secondary institutions must believe that the AP exam is a valid measure. Two inputs that the decision makers at colleges and universities consider are 1) how the test looks and 2) how they believe the AP grades reflect performance in the corresponding course at their institution.

The first criterion, how the test looks, is a major factor in deciding whether to accept AP for credit or placement. We might refer to this as content and “face validity.” Kane (2006) notes that face validity refers to the apparent relevance of tasks to the proposed interpretation of scores and that while the appearance of relevance does not go far in supporting the appropriateness of interpretation, the lack of such relevance can lend credibility to certain challenges to the intended inferences. Performance on the MC sections of the AP exams counts for a significant portion of the AP final score, but it is primarily the CR items that are most valued and are most noticeable to the teachers, students, and administrators. First, all CR items, rubrics, and sample responses at each score point that are included on the US main form and the international main form if applicable (14 subjects) are released each year after the administration and are available on the AP Central website. Teachers can use these questions in their coursework and students can review them as they prepare for AP exams. Paek et al (2005) reported that most teachers find the disclosed AP topics and rubrics to be useful and that, in general, the more the teacher used these resources throughout the year the better the class performed on the examination.

Many subject matter experts believe that the CR items on AP exams are better able than MC items to test the depth of knowledge and skills that are important to the subject. For example, the skills of historical argumentation and synthesis are relevant to the history subjects. One highly-regarded item type on the history tests is the Document Based Question (DBQ). The DBQ stem contains up to seven historical documents of varying size and the test taker is required to analyze, compare, and synthesize the information into a coherent essay answering one or more questions. In addition to the information provided in the documents the test taker is expected to use additional historical knowledge gained in their class as well as the skills to respond to questions. This item type is consistent with the type of assignment that college students are expected to perform for the introductory classes.

At many institutions for many subjects, in-class tests that are given by college or university faculty require written essays or problem solutions rather than MC responses. This adds to the notion that AP exams should be composed of a significant CR component to provide valid inferences for the intended uses. In addition, many AP stakeholders have strong views regarding the type of CR items that should be included, such as the DBQ described above. For another example, an item requiring a one paragraph response for a history or English exam is perceived to be of little value when contrasted with an item requiring a longer essay response that includes a thesis, supporting arguments and conclusion. These perceptions create parameters for the structure of each AP exam. Simply put, while more measurement precision per unit of testing time might be attained with more MC items and/or shorter CR items, the value of AP depends on the stakeholders' acceptance of the CR items as valid measures of the construct. Our experience working with subject matter experts serving on advisory and test development committees over the past several decades has heightened for us the high regard for CR items on the AP exams. This structural constraint on the test design has implications for the procedures used to support score comparability, which we will discuss below.

A second criterion the higher education stakeholders use to judge the relevance of the AP examinations is how they believe the AP grades will reflect performance in the corresponding course at their institution. This judgment depends on the evaluation of the knowledge and skills required for the examinations as well as the implicit belief that the cutscores are set at the right place and that there is consistency of the AP reported scores regardless of the test form taken.

That is, to be useful to the higher education community, the AP courses and examinations must adhere to curriculum consistent with college courses, measure knowledge and skills commensurate with expectations of college students in those courses, and have a meaningful and consistent score scale over time.

The Accepted Class Evaluation Service (ACES) provided for free to colleges by the College Board may be used to investigate the second criterion at an individual institution using its specific cut-scores and following the performance of AP and non-AP students with regards to subsequent course work and/or grade point average.

Fairness

Fairness to all examinees has many tendrils. For this paper the aspects of interest are differential performance across item types, coachability/test preparation, and security.

Currently Mantel-Haenszel DIF is performed operationally for the MC items on AP exams although several special studies have been performed concerning differential behavior across item types for various demographic groups (Bridgeman and Morgan, 1994; Buck, Kostin, and Morgan, 2002; Mazzeo, Schmitt and Bleistein, 1993).. In addition, the CR items, as well as MC items, undergo Fairness Reviews which ensure that the items do not have content that would be offensive or give an obvious advantage to one group versus another.

Bridgeman and Morgan (1994) reported, for example, that males tend to receive higher scores on MC items while females score relatively higher on CR items. An interesting finding is that students who score relatively higher on MC than CR were about as successful in college courses as students exhibiting the opposite pattern. Other researchers have indicated that item content and topic variability are associated with differential performance by males and females (Buck, Kostin and Morgan, 2002; Mazzeo, Schmitt, and Bleistein, 1993). Breland and Oltman

(2001) reported that males perform better on MC sections but not on the CR sections for Macroeconomics, Microeconomics and Comparative Government and Politics. They reported that a survey of instructors suggested that there were no differences in performance in coursework, however. Breland and Oltman also indicated that students who received course credit due to AP Exam grades did as well or better in higher-level courses in the areas of Macroeconomics and Microeconomics than non-AP students.

As well known, there are appropriate and inappropriate methods of test preparation. Coachability may be less of a concern for AP Exams because of the heavy emphasis on content and skills and the AP course descriptions and the Professional Development opportunities for teachers. That is, there seems to be little to be gained, if anything, to coach a student to the test but everything to be gained to teach the material. That is not to say that using previous test materials to prepare students is not valuable. As previously mentioned, Paek et al (2005) reported that most teachers find the disclosed AP topics and rubrics to be useful and that, in general, the more the teacher used these resources throughout the year the better the class performed on the examination. The emphasis by the teachers is on the skills that are being required in the course and applied on the examination, for example, synthesizing information on the Document Based Question in the histories and the criteria on which the responses are scored.

The CR items are disclosed immediately after the administration and are posted on the College Board web-site along with sample responses and the grading guides. This provides a level playing field for all teachers and students in understanding the item types and the expectations for responding.

Security is an issue that is greatly intertwined with the next topic, equitability. As noted, the CR items are memorable but disclosed every year. This poses a problem for equating which is discussed below. A different issue, however, is pre-knowledge based on time zone differences. Although this has not been a problem for AP at this time there is always need for vigilance especially with the gains in technology such as cell phones and small cameras.

Reliability

As mentioned earlier, the number and type of free-response items differ across the examinations as would be expected. The free-response sections carry 33 to 60 percent weight in the final score in terms of number of points awarded. For example, with a 50-50 weight the free-response items will be weighted such that the highest obtainable score on the free-response is equal to the highest obtainable score on the multiple-choice. The weights are chosen by a committee of subject matter experts to meet their interpretation of the knowledge and skills assessed on each section as well as the match of the item types to authentic work produced by students in college courses. The effect of the weighting schemes on the overall reliability is, of course, of interest to ensure that the students, teachers, and other users are assured that the scores are reliable. Several studies have been performed concerning the weights (see e.g., Hendrickson, Melican and Patterson, 2008) and more studies will be necessary as the new examinations are rolled out and periodically for the on-going examinations. It is frequently noted by psychometricians that the weights are not assigned to optimize reliability but to reflect the subject matter experts' evaluation of the value of the free-response section. Hendrickson, Patterson and Ewing (2010) and Rogers and Swaminathan (2010) present papers in this session that address the weighting and reliability topic.

The exam grades, 1 through 5, are sent to the students and to any eligible institution designated by the student. To be of use, the exam grades provided to the institutions must be reliable. Internal consistency reliability estimates for total raw scores range from .85 to .94 across examinations and estimates of classification accuracy as reported are generally around .90. The reliability estimates for the same examination (e.g., U.S. History) tend to be very close from year to year. So, at the composite level the AP exams would seem to exhibit reasonable reliability estimates. It may be noted, however, the reliability estimates for AP are internal consistency estimates; there are no estimates for parallel forms or test-retest reliability which would be interesting projects to undertake.

Reader[1] reliability studies are periodically performed for each AP exam with a selection of test taker's responses scored independently by two raters. Agreement levels, reliability estimates, and the effect of using rater 1 versus rater 2 on the final assigned scores are computed. Information is used to inform test development practices.

The logistics at an AP reading are of interest. CR items are scored by human readers (i.e., raters). In late spring thousands of college and high school teachers converge on various cities where they spend a week grading the students' responses. Over 10,000 readers scored the 2009 exams, with the number of readers per exam ranging from 35 to 1,100. Reliability of raters' scores, then, is a major contributor to the quality of the final scores reported to students and schools. The sheer logistics and costs of the readings are huge so the feasibility of distributed scoring is being reviewed. The practical implications of a change of this sort on reliability need to be evaluated but as discussed later the possible impact on the AP milieu is equally, if not more, important than costs or small changes in the reliability coefficients. The AP Reading offers educators both significant professional development and the opportunity to network with colleagues. Feedback from readers indicates that many believe this is one of the best professional development opportunities of their career. Many readers serve for multiple consecutive years and

[1] In AP, raters are referred to as readers and in this paper we use raters and readers interchangeably.

the waiting list to become an AP reader is indicative of the positive professional experience afforded. The AP readers at the high school and post-secondary level are powerful proponents of AP in the field because of their in-depth understanding of the large investments of talent, experience, and knowledge required to develop and maintain a program of the caliber of AP.

Prior to the readings, the chief reader and the test developers convene and review hundreds of responses to choose exemplars of each rubric score point and generate prompt specific rubrics and scoring guides to supplement the generic rubrics. In addition, training and verification responses are chosen. At the readings the rubrics and scoring guides and training examples are used to train readers and then a verification round is performed where the readers provide ratings for the verification examples. Readers do not start reading exams operationally until they have reached a level of agreement with verification packets.

During operational readings there are several methods used to evaluate whether readers are using the scoring guidelines consistently and appropriately. The table reader regularly "back reads" responses and compares their (table leader) ratings to those generated by each reader. Any change in accuracy or trends are noted and, if necessary, re-training performed. Further, responses that have been read by the chief reader, question leaders and/or table leaders are salted into the packets given to operational readers and the results are compared. Again, if errors or trends are observed retraining occurs. On only very rare occasions are readers released.

Current research areas include performing generalizability theory studies which could necessitate a different manner in distributing test taker responses to multiple readers. The future may well include electronic scoring (as a double read only) and distributed scoring.

Exam Design

The issues of validity, fairness, reliability and score consistency must be addressed at the inception of a testing program and not be left as statistical evaluations after the fact. The exam design and the equating and linking plans support the interpretations and score comparability claims that can be made. The decisions made about test design and assembly create the foundation of the assessment program. In thinking about requirements, it is helpful to consider both the test form and the specific items. If the same exact form could be administered to all test takers regardless of test date and there were no changes in content that would necessitate updating items, no changes in the way the tests were scored, and no concerns about security or cheating, then scores could be directly compared. Since it's not practical to use the same form repeatedly, variation across forms is controlled by setting test form specifications which include content and skills, structural and statistical aspects and levels of tolerance for variation of specifications across forms. Content and skills specifications include parameters usually in the form of ranges of number or percentages of items that should address specific topics and skills. The level of specificity could be very broad or very detailed. It is necessary to have a large enough pool of items that meet the specifications so that forms can be built as desired.

Structural specifications include requirements for numbers and types of items, test/section/item, and administration directions. Statistical specifications typically include difficulty and discrimination parameter targets at the test form and possibly item level. While content and structural specifications can be met without pretesting of items, data based on examinee or pretest sample performance is typically required for inclusion of statistical specifications. There may be ways to infer the statistical characteristics of an item, but any method that would be used needs to be tried in the application where it would be applied.

For each exam offered by the AP program, the structural specifications include the total testing time, the testing time and number of items for each section (MC and CR), and content coverage allocations. The structure and high-level content specifications for each exam can be found in the Course Descriptions located on the AP Central Website (http://apcentral.collegeboard.com). While the general specifications might include 4 or 5 content areas and percentage allocation, more specific specifications with up to 20 content subareas are often used for assembly by the test developers.

For exams undergoing course and exam review, detailed guidelines for form assembly are being created. These exams also include the structural specifications as noted above albeit at a more refined level. That is, based on input from committees of subject matter experts, much more detail is being introduced and specified targets are being established for multiple factors Hendrickson, Huff and Luecht, 2009). For a history exam these factors could include skills (e.g., crafting historical arguments from historical evidence, chronological reasoning, etc.), themes (e.g., interaction between humans and the environment, development and transformation of social structures, etc.), time periods, geographical region, and key concepts within periods and themes.

There are other design considerations at the item level. For each CR item, item design specifications for prompts and scoring rubrics are included. For example, each of the three AP History Exams and the AP English Language and Literature Exam include a Document Based Question (DBQ) and the test specifications include the number and types of reference documents for the question. Across the spectrum of AP Exams, types of scoring rubrics vary generally along the analytical to holistic continuum. However for each exam, the types of rubrics are consistent from form to form. For example, the AP Calculus Exams have analytical rubrics, while the two

AP English Exams have three essay questions scored with holistic rubrics that usually have a 0-9 point possible score range. Within an Evidence-Centered Design process, the guidelines might also include sampling and use of specific task models by item writers to generate items (see Hendrickson, Huff & Luecht, 2009 for a description of task models).

With respect to AP, some unique practical issues related to fairness and acceptance by users of the test design must be addressed. For each AP subject, there are course outlines and curricular requirements (see AP Course Audit and Course Descriptions which can be found under this page http://www.collegeboard.com/html/apcourseaudit/index.html). More detailed curricular frameworks are being developed for each course as it undergoes course and exam review.

The independence of individual teachers to have flexibility in how they structure and teach their AP courses is a highly valued and important mainstay of the AP Program. For example, each of the AP history courses, and AP World History in particular, continues to span an enormous amount of content, not all of which can or needs to be studied in depth. Required content for both the course and the exam include all Historical Thinking Skills and all elements of the Curriculum Framework’s content outline. Teachers will, and should, however, vary the depth to which they explore the various topics.

Given the variability among the AP history courses, the questions on the exam need to test in depth the historical thinking skills and abilities of all students while having the question topic fair to students regardless of their teachers’ choice of which topics to cover in more depth than others. For MC questions, decisions about the cutscore points must take this issue under consideration. For CR questions which are time consuming and heavily weighted however, a different approach is needed.

One way that this has been addressed in some AP exams to date has been by offering choice of topic to candidates. The following examples can be found by looking at the CR items posted on the AP Central website and accessed through this link: http://apcentral.collegeboard.com/apc/public/courses/teachers_corner/index.html. This can be via external choice of two or three different questions (US History and European History essay questions 2 and 3), internal choice of a limited number of options (World History essay question 3–2002 through 2008, and Latin Vergil question 5–2009), or internal choice of an unlimited nature (English Literature essay question 3). From a psychometric stance, choice can introduce noise into the measurement and may not work as well as intended for the test takers.

Some test takers do not choose the prompt that would allow them to get the higher score. Wang, Wainer and Thissen (1995) performed a small study regarding choice using MC Chemistry items. They required test takers to indicate a preference for one of two multiple-choice items but answer both items using three pairs of items. Wang et al. reported that for one pair of items the set of test takers who chose item 12 did far better on item 11, the non-choice items. Thus, there test takers chose incorrectly and would have disadvantaged themselves. Results were less pronounced for the other two pairs of items.

Bridgeman, Morgan and Wang (1997) performed a similar experiment using CR items in classrooms prior to the national examination. Two subjects, AP U.S. History and AP European History were involved. The means for the preferred topic were greater than the scores of the prompt that was indicated as the one that was not preferred and would not have been chosen. The percent of students with lower scores on the preferred topic, however, varied from 29 to 32 percent. Bridgeman, et al did stress that failure to allow choice could disadvantage students unfamiliar or uncomfortable with a topic when no choice is allowed. They ended with recommendation for trying to keep difficulty of topics approximately equal and equating scores from the subforms created by allowing choice. The AP program has embarked on an effort to use Evidence Centered Design (ECD) (Ewing, Packman, Hamen and Clark, 2009; Hendrickson, Huff and Luecht, 2009; Huff and Matts, 2009;and Mislevy, Almond, and Lukas, 2003) to develop test specifications, test design, and item generation which may help address some of the concerns raised by Bridgeman et al.

From a practical standpoint, subject matter experts are often in favor of choice. The arguments people provide for wanting choice are, in fact, not unreasonable. Opportunity to learn is an issue if the content of the item has been thoroughly treated by one teacher and not another, for example. Thus, allowing a student to choose a topic containing content that is familiar is a seemingly reasonable option.

External choice as practiced today is most problematic because of the lack of parallelism and equitability. That is, not only is content apparently different but the skills, level of skills, or interplay of skills with content can vary considerably, even when not intended by item writers to do so. In addition, some test takers choose badly (Wang, Wainer and Thissen, 1995; Bridgeman, Morgan and Wang, 1997) and disciplined approach to the exam design such as ECD to generate alternative prompts that are measuring the same skills but allowing different content in one of the two following ways might be the most feasible compromise. First if the question is designed to test skills and the rubrics are written to address the skills component then internal choice with an unlimited or wide selection of content topics would be preferred. Second, the use of a structured item specification framework and task models for the CR items would be helpful when limited internal choice or external choice is desired. Thus a set of items that are in a limited selection internal choice or an external choice grouping could be specified to test the same skills, be generated from tightly constrained task models, and be scored against the same rubrics.

The effect on test reliability and consistency with this type of solution is yet unknown. However it seems a reasonable attempt to address the root concern of the SMEs with the best chance at fairness to all test takers getting the specific form. Once within-form structural parameters are defined, the more they can be applied across forms, the more comparable and equitable the scores on different forms are likely to be.

In deciding on the level of design specificity that can be controlled across forms of the test, practical considerations will need to be balanced with the desired ideal. The practical considerations include availability of pretesting, administration conditions, cost constraints, timelines and perception of face validity by users. It is advisable and best practice to design as much control and consistency in the test development and assembly process as feasible to assure comparability across forms and fairness to examinees.

Equitability

Along with well planned and specified test development procedures, the equating and linking plan contributes to the score comparability across exam administrations and forms. There are a number of decisions to be made as part of this planning, many of which will be subject to practical constraints. Some of the practical constraints faced by the AP program were described above.

Four high-level considerations for equating and linking plans include: 1) What are the form use needs? 2) What data will be available to use for equating and linking? 3) What options are there for reuse of MC and CR items? and 4) If cutscores are required, are there any additional steps needed to assure consistency of cutscores across forms?

A primary consideration is the need for forms or item banks. In the simplest case only one form would be needed and all examinees could be administered that form regardless of test date or location. For most programs, including AP, that would be untenable and multiple forms are needed for different testing locations and test dates. While beyond the scope of this paper, it should be noted that many testing programs utilize banks of items and then generate forms based on assembly specifications, sometimes as examinees are testing. This has been done for a number of computer administered exams. While these programs don't create a small number of fixed forms, they take into account specifications for active item banks and rotation of items in addition to form assembly algorithms.

The volume and time availability of data for equating inform the equating design plans. The AP exams vary greatly in volume as well as test taker population characteristics. First, it is important to meet minimum volume requirements to run analyses of interest. Equally important is that those analyses are conducted on the appropriate and representative samples. Most large scale testing programs face tight score reporting deadlines that affect all stages of processing, and AP is no exception.

Many of the AP exams are administered to over 100,000 examinees so minimum volume requirements are not a concern. Other AP exams are administered to fewer than 5,000 examinees annually. For both the small and large volume exams, conducting the equating analyses on a good sample given the processing time constraints is a concern. For example, AP World Language courses are targeted to non-heritage speakers, and the exam is calibrated such that successful performance of AP students on the exam is intended to be equivalent to the expected performance of students who have completed three years (five or six semesters) of college language courses at postsecondary institutions. Many heritage speakers who have spoken the language since birth sit for AP World Language exams and their performance is not necessarily representative of the intended AP test taker population. Sampling issues such as this must be considered in conducting the equating analyses.

The decision about CR item reuse also provides a constraint for designing a linking plan. At one extreme, CR items can be used and reused freely on any form (unlimited and unconstrained reuse) and at the other extreme CR items can be used once and only once (no reuse).

Taking these considerations into account, the scheme for equating and linking can be created. Two widely used methods for equating and linking scores on different test forms are the common item anchor test design (NEAT) and the equivalent groups design. Holland and Dorans (2006) and Kolen and Brennan (2004) provide thorough descriptions of these procedures. In the equivalent groups design, a spiraling process is used to randomly assign forms to examinees and the result is that the groups of examinees taking each form are randomly equivalent. Their performance data are then used to link the scores on different forms. In the anchor test design, test forms have a set of items in common and different groups of examinees are administered the two forms. The administrations in this case are usually on separate dates. For the results to be most accurate, the set of common items should be proportionally representative of the total test form in content and statistical characteristics.

With mixed format tests, if there is no or limited reuse of CR items, then an anchor test design might not be fully supported. As mentioned above, for a number of reasons the CR items on AP exams are not routinely reused and cannot be included in an anchor test. With only MC items in the anchor set, the quality of the anchor test equating likely varies across AP exams. One of the papers in this session looks specifically at this issue in the context of AP (Lee, Kolen, Hagge & He, 2010). Another paper in this session compared non-linear equating methods given a common item set with only MC anchors (Powers, Liu, Hagge and He, 2010). Also of note here is a paper in this session which uses IRT mixture modeling to detect unobserved subgroup differences on mixed format exams (Kaliski and Barry, 2010).

Recently there has been thought given to including CR items in the anchor set for AP Exams. First, the security concerns need to be adequately addressed. Beyond that, depending on the test characteristics the inclusion of CR items in the anchor set may or may not improve the quality of the anchor test equating. Within some tests the CR items vary such that it would be difficult to select a subset of the CR items to serve as a mini test that adequately represents the CR section.

With regard to maintaining cutscores, for a testing program that categorizes scores into pass/fail or multiple categories such as the AP scores of 1 through 5, once initial cutscores are set on a reference form ideally they can be applied to other exams via the equating and linking process. However, it is worth noting here that in some cases where the feasible equating procedures may contain more equating error than deemed acceptable, additional policy review may be indicated before scores are finalized.

Conclusion

Mixed-format examinations such as those used by the AP Program must provide valid and fair interpretations for all subgroups across years for the exam scores to be useful to the user communities. This is, of course, do-able as testing programs have been meeting these goals for many years. The issues of validity, fairness, reliability and score consistency can be addressed but for mixed-format tests there are many decisions to be made and no examination or examination program faces exactly the same choices. This paper raised some issues and used the

AP experience to illustrate the types of questions that need to be addressed and outlined the strong test development processes that need to be in place to address the issues.

References

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (ETS RR-90-7). Princeton, NJ: Educational Testing Service.

Breland, H.M. and Oltman, P.K. (2001). *An Analysis of Advanced Placement (AP) Examinations in Economics and Comparative Government and Politics.* (College Board Research Report No. 2001-4). New York: The College Board.

Bridgeman, B. and Morgan, R. (1994) *Relationships between Differential Performance on Multiple-choice and Essay Sections of Selected AP ® Exams and Measures of Performance in High School and College.* (College Board Research Report No. 1994-5) New York: The College Board.

Bridgeman, B., Morgan, R., & Wang, M-M. (1997). Choice Among Essay Topics: Impact on Performance and Validity. *Journal of Educational Measurement, Volume 34, Number 3,* 273-286.

Buck, G., Kostin, I., & Morgan, R. (2002). *Examining the relationship of content to gender-based performance differences in Advanced Placement Exams.* (College Board Research Report No. 2002-12). New York: The College Board.

Dodd, B. G., Fitzpatrick, S. J., De Ayala, R. J., & Jennings, J. A. (2002). *An investigation of the validity of AP grades of 3 and a comparison of AP and non-AP student groups.* (College Board Research Report No. 2002-9). New York: The College Board.

Dougherty, C., Mellor, L., & Jian, S. (2006). *The relationship between Advanced Placement and college graduation.* National Center for Educational Accountability: 2005. (AP Study Series, Report 1). Austin, Texas: National Center for Educational Accountability.

Ewing, M. (November 2006) *The AP® Program and Student Outcomes: A Summary of Research* (College Board Research Note, RN-29). New York: The College Board.

Ewing, M., Packman, S., Hamen, C., and Clark, A. (2009). *Representing Targets of Measurement Using ECD.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Hargrove, L., Godin, D. and Dodd, B. (2008). *College Outcomes Comparisons by AP and Non-AP High School Experiences.* (College Board Research Report No. 2002-4) New York: The College Board.

Hendrickson, A., Huff, K., and Luecht, R. (2009). *Claims, Evidence and Achievement Level Descriptors as a Foundation for Item Design and Test Specifications*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Hendrickson, A., Melican, G., and Patterson, B. (2008) *The Effect of Using Different Weights for Multiple-Choice and Free-Response Item Sections*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Hendrickson, A., Patterson, B., and Ewing, M. (2010). *Developing Form Assembly Specifications for Exams with Multiple Choice and Constructed Response Item Section*. Paper presented in Poster Session, Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple Choice and Constructed Response Items at annual meeting of the National Council on Measurement in Education, Denver.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan, (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.

Hogan, T.P. and Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What experts say. *Applied Measurement in Education, Volume 20, Number 4,* 427-441.

Huff, K., Steinberg, L., and Matts, T. (2009). *The Promise and Challenge of Implementing ECD in Large Scale Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Kaliski, P. and Barry, C. (2010). *Using Mixture Modeling to Detect Unobserved Subgroups on Standardized Mixed-Format Exams.* Paper presented in Poster Session, Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple Choice and Constructed Response Items at annual meeting of the National Council on Measurement in Education, Denver.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practice* (2nd ed.). New York, Springer.

Lane, S. and Stone, C.A. (2006) Performance Testing. In R.L. Brennan, (Ed.) *Educational Measurement,* (4th edition, pp 387-432). Westport, CT: American Council on Education/Praeger.

Lee, W.C., Kolen, M.J., Hagge, S. & He, Y.H. (2010*). Equating Mixed Format Tests Using Dichotomous Anchor Items*. Paper presented in Poster Session, Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple Choice and Constructed Response Items at annual meeting of the National Council on Measurement in Education, Denver.

Mazzeo, J., Schmitt, A.P., and Bleistein, C. A. (1993). *Sex-Related Performance Differences on Constructed-Response and Multiple-Choice Sections of Advanced Placement Examinations*. Educational Testing Service Research Report, RR-93-05. ETS, Princeton, NJ.

Mislevy, R.J., Almond, R.G., and Lukas, J.F. (2003) *A Brief Introduction to Evidence Centered Design.* Educational Testing Service Research Report, RR-03-16. ETS, Princeton, NJ.

Paek, P.L., Ponte, E., Sigel, I., Braun, H., and Powers, D. (2005). *A Portrait of Advanced Placement Teachers' Practices*. (College Board Research Report No. 2005-7). New York: The College Board.

Powers, S., Liu, C., Hagge, S., and He., Y. (2010). *A Comparison of Nonlinear equating Methods for Mixed Format Exams.* Paper presented in Poster Session, Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple Choice and Constructed Response Items at annual meeting of the National Council on Measurement in Education, Denver.

Rogers, J. and Swaminathan, H. (2010). **S***coring and Combining Multiple Choice and Free Response Items in Mixed Format Tests.* Paper presented in Poster Session, Mixed-Format Tests: Addressing Test Design and Psychometric Issues in Tests with Multiple Choice and Constructed Response Items at annual meeting of the National Council on Measurement in Education, Denver.

Sadler, P.M., Sonnert, G., Tai, R.H., and Klopfenstein, K. (2010). *AP: A Critical Examination of the Advanced Placement Program.* Harvard Education Press

Wang, X.B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, Volume 8, Number 3,* 211-226.

Welch, C. (2006). Items and prompt development in performance testing. In T.M. Haladyna, and S.M. Downing, (Eds.) *Handbook of Test Development*, (pp. 303-327). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Table 1: Features and 2009 Volumes of AP Exams

Subject	2009 Volume	Special Features	Exam Design CR
Arts			
Art History	20,619	Until 2009 Slides (Some MC & CR); 2010 printed inserted	Essay; short essays with picture prompts
Music	15,438	Listening & Sight Singing (Recorded)	Music writing: dictation with listening prompts; sight singing
Studio Art- Drawing	14,589	Portfolio	3 sections: Quality (artwork submitted); Concentration and Breadth (digital submission)
Studio Art -2D Design	17,387	Portfolio	3 sections: Quality (artwork submitted); Concentration and Breadth (digital submission)
Studio Art - 3D Design	2,761	Portfolio	3 sections: Quality, Concentration and Breadth (digital submission)
English			
English Literature	332,352	Internal choice, unlimited	3 Essays
English Language	337,441	1 Synthesis essay	3 Essays
World Languages			
Chinese Language & Culture	5,100	Listening & Speaking; Computer Delivery	Presentational & interpersonal speaking and writing
French Language	21,029	Listening & Speaking	Writing and speaking
German Language	5,001	Listening & Speaking	Writing and speaking
Japanese Language & Culture	2,085	Listening & Speaking; Computer Delivery	Presentational & interpersonal speaking and writing
Latin Vergil	4,295	Internal choice (limited)	2 translations, 1 long essay, 2 short essays
Spanish Language	110,723	Listening & Speaking	Presentational & interpersonal speaking and writing
Spanish Literature	16,633	Internal choice (limited to 2); Content & Lang scores on essay	3 Essays

Continued on next page

Table 1 (cont.): Features and 2009 Volumes of AP Exams

Subject	2009 Volume	Special Features	Exam Design CR
History			
European History	101,359	External choice essays	DBQ + 2 Essays
US History	360,173	External choice essays	DBQ + 2 Essays
World History	143,426	Internal choice essays (Limited to 2 or 3)	DBQ + 2 Essays
Math			
Calculus AB	230,588	Calculator (Some MC & CR)	Problem sets
Calculus BC	72,965	Calculator (Some MC & CR)	Problem sets
Computer Science A	16,622		Program coding
Statistics	116,876	Calculator	Short answer sets; 1 investigative task
Science			
Biology	159,580		Set-based short answer
Chemistry	104,789	Calculator (Some CR)	Problem-based sets; essay/short answer
Environmental Science	73,575		Set-based short answer
Physics B	62,702	Calculator for CR	Problem-based sets
Physics C: Mechanics	29,167	Calculator for CR	Problem-based sets
Physics C: Elec & Mag	12,628	Calculator for CR	Problem-based sets
Social Science			
Microeconomics	46,272		Sets with graph creation, short answer and short essay
Macroeconomics	73,817		Sets with graph creation, short answer and short essay
Comp Gov & Politics	14,728		Short essay/short answer sets
US Gov & Politics	189,998		Short essay/short answer sets
Human Geography	50,730		3 Short answer/short essay sets
Psychology	151,006		2 Essays