Available online at www.sciencedirect.com
SCIENCE DIRECT®
Contemporary Educational Psychology 31 (2006) 344–376
Contemporary Educational Psychology
www.elsevier.com/locate/cedpsych

Using the theory of successful intelligence as a basis for augmenting AP exams in Psychology and Statistics

Steven E. Stemler [a,*], Elena L. Grigorenko [b], Linda Jarvin [c], Robert J. Sternberg [c]

[a] *Department of Psychology, Wesleyan University, 207 High Street, Middletown, CT 06459, USA*
[b] *Department of Psychology and Child Study Center, 2 Hillhouse Avenue, New Haven, CT 06520, USA*
[c] *Tufts University, Ballou Hall, Medford, MA 02155, USA*

Available online 23 January 2006

Abstract

Sternberg's theory of successful intelligence was used to create augmented exams in Advanced Placement Psychology and Statistics. Participants included 1895 high school students from 19 states and 56 schools throughout the U.S. The psychometric results support the validity of creating examinations that assess memory, analytical, creative, and practical skills in the context of content-specific knowledge. In addition, Q-factor analyses revealed a set of empirically distinguishable profiles of achievement, supporting the assertion that individuals exhibit different patterns of strengths and weaknesses in cognitive processing skills. Finally, an examination of ethnic group differences in achievement shows that measuring a broad range of cognitive skills tends to reduce ethnic differences in achievement. Future studies aimed at replicating these findings are warranted.

Keywords: Advanced placement exam; Statistics; Psychology; Cognitive skills; Theory of successful intelligence; Learning; Assessment; Ethnic group differences; Cognitive profiles; Achievement testing; Q-factor analysis; Cluster analysis

* Corresponding author. Fax: +1 860 685 2761.
E-mail address: sstemler@wesleyan.edu (S.E. Stemler).

0361-476X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cedpsych.2005.11.001

1. Introduction

Each year, millions of students across the country take high-stakes achievement tests that will have an important influence on their academic and professional future (Heubert & Hauser, 1999); yet, many of these tests are not aligned with modern theories of student learning and cognitive processing. As a result, students with strengths in cognitive skills not assessed by these tests may have their future opportunities curtailed (Sternberg, 1997). For example, many students with strong creative or practical skills but weaker memory and analytical skills never have the opportunity to reach the highest levels of education, where they might thrive, because the tests that are used as gatekeepers tend to emphasize a more limited range of skills (e.g., memory and analytical skills) than might be optimal. Yet, a narrow range of skills, such as memory and analytical skills, taken alone is not sufficient to succeed in the professional world. Instead, a balance of a wider range of cognitive skills is important, regardless of one's professional domain.

Broadening the range of cognitive skills assessed is important not only at the individual level. It has potentially important implications at the group level as well. Sternberg and colleagues (Sternberg and The Rainbow Project Collaborators, 2005b; Sternberg, Torff, & Grigorenko, 1998a,1998b; Sternberg et al., 2004) have shown that when assessments are designed to measure a broad range of cognitive skills, the achievement gap typically observed between White students and underrepresented minority students (Chubb & Loveless, 2002; Jencks & Phillips, 1998) appears to be reduced substantially.

In recent years, designers of large-scale testing programs, recognizing the important social, economic, and ethical consequences associated with standardized testing, have become increasingly interested in linking educational assessment to modern theories of cognitive processing skills (Embretson & Reise, 2000; Irvine & Kyllonen, 2002). To the extent that high-stakes exams draw on sound traditions of research in psychological theory and educational assessment, the results will be more construct valid and defensible.

Therefore, in the spirit of infusing cognitive theory into educational assessment, the aim of the current study was to create a series of augmented exams for the College Board's Advanced Placement (AP) program that would be explicitly based on one validated theory of cognitive processing (Sternberg, 1985, 1997, 1999), the theory of successful intelligence. We were particularly interested in examining individual and ethnic group differences in cognitive processing skills within the context of AP Psychology and Statistics.

Of course, this is not the only theory that could serve as a basis for such an assessment. There are many others (Alexander, Jetton, & Kulikowich, 1995; Carroll, 1993; Cattell, 1971; Ceci, 1996; Gardner, 1983; Luria, 1973) that might also serve as a basis for augmentation of existing tests. Perhaps future studies will compare alternative theories as a basis for such augmentation. We chose the theory of successful intelligence, in particular, because (a) it has been validated through converging operations in a number of different studies, (b) it has rather clear implications for operationalization in the context of item construction for AP exams, (c) past studies had shown incremental validity for theory in the context of assessment, and (d) we are familiar with the theory and its implications.

2. Background

To set the context for the study, we begin with a brief description of the AP program. We then present our theoretical framework for the study and briefly review the literature related

to individual differences in cognitive skills and ethnic differences in achievement. This section concludes with a statement of the specific research questions under investigation.

2.1. The Advanced Placement program

The College Board's Advanced Placement program, started in 1955, was originally designed as a mechanism for granting exceptional high school students the opportunity for advanced study that would be equivalent to college-level programming. Since then, the program has expanded both in terms of the kinds of students eligible to take the courses and in the number of different subject areas covered by the program. Over time, the program has become widely disseminated. In 2002, a total of 937,951 students (about 10% of all high school students) took an exam in one of the 34 courses across 19 subject areas offered by the AP program.[1]

Each spring, students enrolled in AP courses are given the opportunity to take a high-stakes exam to demonstrate their mastery of the subject area. The exams are graded on a scale from 1 to 5, "with five indicating a student who is extremely well-qualified to receive college credit and/or advanced placement based on an AP exam grade" (College Board, 2004). Typically, students scoring 3 or higher on the exam are eligible in many colleges to receive college credit for their course. Thus, the results of the test have potentially important financial implications, as placing out of the college courses potentially can save a student thousands of dollars in tuition in subsequent years. The limited number of chances to take the test, coupled with the potentially significant financial savings associated with the outcome, qualifies the AP exam as a high-stakes test.

Traditionally, the chief concern of the AP exam developers has been with the assessment of particular subdomains of academic content and skills rather than with the explicit assessment of students' cognitive skills. Given the high consequences attached to the test results of AP exams, the designers of the AP program are now seeking to ensure that its tests are aligned with the latest thinking about how students learn.

2.2. Theoretical framework

According to Sternberg's theory of successful intelligence (1984, 1985, 1997, 1999), a common set of processes underlies all aspects of problem solving. These processes, although not their behavioral manifestations, are hypothesized to be universal (Sternberg, 2003). For example, although the solutions to problems that are considered intelligent in one culture may be different from the solutions considered to be intelligent in another culture, the need to define problems and translate strategies to solve these problems exists in any culture. *Metacomponents*, or executive processes, plan what to do, monitor things as they are being

[1] The 34 courses offered by the AP program include: Art History, Biology, Calculus AB, Calculus BC, Chemistry, Computer Science A, Computer Science AB, Economics: Macro, Economics: Micro, English Language and Composition, English Literature and Composition, Environmental Science, European History, French Language, French Literature, German Language, Government and Politics: Comparative, Government and Politics: United States, Human Geography, International English Language/APIEL, Latin Literature, Latin: Vergil, Music Theory, Physics B, Physics C: Electricity and Magnetism, Physics C: Mechanics, Psychology, Spanish Language, Spanish Literature, Statistics, Studio Art: 2-D Design, Studio Art: 3-D Design, Studio Art: Drawing, U.S. History, World History.

done, and evaluate things after they are done. *Performance components* execute the instructions of the metacomponents. *Knowledge-acquisition components* are used to learn how to solve problems or simply to acquire declarative knowledge in the first place.

Although the same components are used for all three aspects of intelligence universally, these processes are applied to different kinds of tasks and situations, depending on whether a given problem requires analytical thinking, creative thinking, practical thinking, or a combination of these kinds of thinking. In particular, *analytical* thinking is invoked when components are applied to fairly familiar kinds of problems abstracted from everyday life. *Creative* thinking is invoked when the components are applied to relatively novel kinds of tasks or situations. *Practical* thinking is invoked when the components are applied to experience to adapt to, shape, and select environments. Thus, the same components, applied in different contexts, yield different kinds of thinking—analytical, creative, and practical. Ultimately, one needs creative thinking to generate new ideas, analytical thinking to determine if they are good ideas, and practical thinking to implement the ideas and to persuade others of their value.

The theory of successful intelligence is not wholly incompatible with aspects of other theories, such as Bloom's (1956) taxonomy of cognitive skills and Gardner's (1983) theory of multiple intelligences. Ultimately, however, the usefulness of any one theory for augmentation of a test such as the AP exam is shown by empirical data examining what happens when the test is augmented by the particular theory.

A key advantage to using an expanded theory of cognitive-processing skills in test construction is that it can provide useful information about and for individual students. Within the theoretical framework of the theory of successful intelligence, students could receive a score report showing their specific profile of strengths and weaknesses across a variety of cognitive skills, which they then could use in future learning opportunities to capitalize on their strengths and compensate or correct for their weaknesses. Furthermore, by measuring a broader range of cognitive skills, individuals who might have been labeled as low achievers when assessed on a limited set of cognitive skills may have better opportunities to demonstrate their content area mastery.

2.3. Differences in student achievement

One of the biggest challenges facing the AP program is in the recruitment of minority students to participate in the program. In 2002, approximately 14% of all students who took one or more exams were African American or Latino, a figure substantially lower than their relative representation in the high school population of 30%.[2] The demographic breakdown of participants varies some by subject area. For example, in 2002, 70% of test takers in AP Psychology were White students, 5% were African American, and 7% were Latino. In AP Statistics, 68% of the test-takers were White, 4% were African American, and 6% were Latino.

In addition to the problem of low minority student enrollment in advanced courses, one of the most persistent problems in instruction and assessment over the years has been the existence of systematic differences in student achievement by ethnicity. Many authors have noted the persistent presence of a Black–White test score gap (first documented in 1966), with White students tending to outperform minority students on most conventional tests of

[2] African-American and Latino students represented 30% of the secondary school population in 1996, the last year this information was collected by NCES's Youth survey, and their numbers have continued to grow.

achievement by nearly a full standard deviation (Chubb & Loveless, 2002; Jencks & Phillips, 1998). Researchers have proposed several possible reasons for these results, including genetic differences (Herrnstein & Murray, 1994), cultural differences (Fordham & Ogbu, 1986; Williams, 2004), and social psychological differences (Steele, 1997). We believe that one reason for this persistent difference is that traditional achievement tests tend to assess a fairly limited range of cognitive skills, ignoring other important skills.

Sternberg and colleagues (Sternberg & The Rainbow Project Collaborators, 2005b; Sternberg et al., 1998a, 1998b) have shown in a series of studies that when assessments are designed to expand the range of cognitive skills assessed, the achievement gap between White students and minority students can be reduced. For example, in a recent study designed to create assessments that would augment the predictive power of the SAT, Sternberg and the Rainbow Project collaborators found that adding assessments of creative and practical skills roughly doubled the power to predict first-year-college GPA compared with the use of the SAT alone. Furthermore, differences in achievement between White students and Black students were drastically reduced (typically by about 0.5 *SD*) on measures of creative skills as compared with assessments emphasizing analytical skills (Sternberg & The Rainbow Project Collaborators, 2005a, 2005b). Similarly, differences between White students and Latino students were reduced on assessments emphasizing practical skills and creative skills as compared with analytical skills (typically about 0.8 *SD* on both creative and practical assessments). Thus, it appears that not only do individual differences in profiles of strengths and weaknesses exist across cognitive skills, but there are also systematic group differences as well.

2.4. Research questions and hypotheses

The purpose of this study was to examine individual and group differences in achievement with respect to cognitive-processing skills within the contexts of AP Psychology and Statistics. In particular, we were interested in examining the following research questions:

1. Is it possible to develop psychometrically sound assessments based on the theory of successful intelligence in the context of AP Psychology and Statistics?
2. Do students, in general, show uneven profiles of strengths and weaknesses across different cognitive skills or do students generally exhibit a relatively even profile of strengths and weaknesses across cognitive skills?
3. Are there systematic ethnic-group differences in achievement across different cognitive-skill areas, regardless of the content domain assessed?

In our study, we hypothesized that the analytical and memory subscale scores from the augmented exams would exhibit the strongest relationship with the scores from the actual AP exam. This hypothesis was based on the view that traditional tests are strongest at measuring memory and analytical rather than creative and practical skills.

3. Methods

In this section, we describe the process by which the AP exams in Psychology and Statistics used in this study were developed. We then describe how each exam used in this study compared with the actual AP exam in that same subject area. This section then concludes with a description of the sample of students and teachers used in this study.

3.1. Instruments

To investigate our research questions, it was necessary to develop alternative versions of the AP exams in Psychology and Statistics. These "augmented" exams were designed to mimic the actual AP exams as much as possible; however, the augmented exams were also developed with an eye toward explicitly balancing items for the cognitive skills they assessed.

The newly developed items for both exams were systematically designed to follow a particular structure. Memory-based items tended to ask the respondent to recall or recognize simple factual information. The stem often provided direct cues or asked for definitions. Items designed to tap analytical skills tended to ask respondents to compare and contrast, critique, evaluate, or judge something. (Memory is not a separate part of the theory of successful intelligence, but rather, is important in all of its other parts. Analytical, creative, and practical processing all operate on information stored in long-term memory and working memory. Moreover, traditional tests emphasize memory for information, and hence it was important that our augmented exams, like traditional exams, include items assessing primarily memory-based performance.) Analytical items typically dealt with abstract and academic, rather than concrete or practical, concepts. They required participants to analyze, evaluate, critique, or compare and contrast. Creative items required the respondent to imagine, suppose, discover, or invent. Creative items often involved a novel analogy, a low-probability situation, or a suspension of conventional beliefs. Practical items required the respondent to apply, use, or implement a concept within a social context. The stem of a practical item typically presented the respondent with a goal or a context for solving the problem. The response options for multiple-choice items often included the application of a concept rather than a naming of the concept. In the next session, we present some example items that were designed to tap each of the aforementioned processing skills.

3.1.1. Assessing memory, analytical, creative, and practical abilities: Example items

Memory items require students to recall and/or recognize who did certain things (e.g., proposed a theory), what things they did (e.g., the nature of the theory), how certain things are done (e.g., computing a standard deviation), when certain things are done (e.g., when squaring of terms is done in a formula), etc.

For example, consider the following multiple-choice question:

According to the psychologist Carl Rogers, which are the three conditions for promoting human growth and fulfillment?

(a) gentleness, kindness, and empathy
(b) genuineness, acceptance, and empathy
(c) agreeableness, acceptance, and extroversion
(d) genuineness, generosity, and extroversion
(e) creativity, generosity, and empathy

The correct answer to this question (underlined) requires knowledge of the theory of Carl Rogers. The student needs to rely in part on his or her long-term memory, answering this question. However, the question can also be answered through a combination of memory and analytical abilities: if the participant understands the theory of Rogers, he or she can infer what the three conditions are most likely to be. Indeed, many memory items

require, or can be solved through, the use of at least some inference. Items were classified as "memory" if they were adjudged to require primarily memory for correct solution.

Analytical items require students to analyze (e.g., Freud's theory of depression), critique (e.g., the design of an experiment), evaluate (e.g., whether a certain formula is appropriate for solving a statistical problem), compare and contrast (e.g., two statistical tests of significance), and so on.

For example, consider the following multiple-choice question:

Suppose you earned 75 points on the most recent exam in statistics. The teacher announced that the mean score for the class was 87 points with a variance of 27.04. What can you conclude about your grade in relation to that of your peers?

(a) Your performance was slightly lower than that of the rest of the class.
(b) Your grade is substantially lower than that of the rest of the class.
(c) Your grade is higher than the majority of your classmates.
(d) Statistically speaking, your grade is about the same as the rest of the class.
(e) There is not enough information given to allow make any conclusions regarding your grade in relation to the rest of the class.

In finding the correct answer (underlined), the student is expected to rely on his or her understanding of the normal curve to analyze the proposed choices. Unlike the memory item above, this item *cannot* be solved purely by memory.

Creative items require students to create (e.g., the design of an experiment), imagine (e.g., how a theory of intelligence would apply cross-culturally), invent (e.g., a theory), or suppose (e.g., what would happen if an achievement test designed for American children was translated and then administered to children in rural Kenya).

For example, consider the following multiple-choice question:

Imagine that you had to produce a TV sitcom to illustrate Freud's personality theory. Which of the following characters would best represent the superego?

(a) A firefighter
(b) An action-movie hero
(c) A nurse
(d) An artist
(e) A Supreme Court judge

Answering this question, the student is expected to imagine the described theory and map it onto the offered selection of answer options.

Practical items require students to apply (e.g., the formula for conducting an independent-samples *t* test to an everyday problem involving comparing prices of two brands of gasoline measured across various service stations), use (e.g., a theory of dreaming to understand why someone had a certain dream), apply what has been learned (e.g., the difference between the mean, median, and mode to deciding which statistic should be used in computing average incomes in a highly right-skewed sample of incomes).

For example, consider the following item:

By mowing your neighbor's lawn for pay, you started earning your own money in 1994. Since then, your personal income has grown every year. Your best summer was the summer of 1997—you had a great job and made some money. You decided to analyze

the dynamics of your income over the 7 years (1994–2000) and fitted a least-squares regression line to these data. Then you decided to recode the data so that the year 1997 was labeled as 0. Now the years are coded by $\{-3, -2, -1, 0, 1, 2, 3\}$. Using these coded data, you fitted another least squares regression line. Compare the slope and intercept of the newly fitted regression line to those of the original regression line. Which of the following is true?

(a) Slope stayed the same, intercept decreased.
(b) Slope stayed the same, intercept increased.
(c) Slope increased, intercept increased.
(d) Slope decreased, intercept increased.
(e) No change of slope or intercept.

In answering this item, the students are expected to activate their knowledge of Exploratory Data Analysis and apply their knowledge in the context specified by the stem above.

In addition to assessing these cognitive skills using multiple-choice types of items, we assessed the skills with a series of open-response items. A single item could have up to four subcomponents, each relating to a different processing skill, as in the following examples:

A variety of explanations have been proposed to account for why people sleep.

(a) Describe the Restorative Theory of sleep.
(b) An alternative theory is an evolutionary theory of sleep, sometimes referred to as the "Preservation and Protection" theory. Describe this theory and compare and contrast it with the Restorative Theory. State what you see as the two strong points and two weak points of this theory compared to the Restorative Theory.
(c) How might you design an experiment to test the Restorative Theory of sleep? Briefly describe the experiment, including the participants, materials, procedures, and design.
(d) A friend informs you that she is having trouble sleeping. Based on your knowledge of sleep, what kinds of helpful (and health-promoting) suggestions might you giver her to help her fall asleep at night?

Part (a) would be an example of an item primarily requiring memory abilities. Parts (b)–(d) would be examples of items primarily requiring analytical, creative, and practical abilities, respectively. Another example open-ended item is presented below, this time from the domain of statistics:

A manufacturer claims that under typical road-travel conditions, the wear of tire tread after 50,000 miles for the manufacturer's tire is approximately normally distributed with a mean of 2 mm and *SD* of 0.2 mm. A tire is determined to be unsafe if the wear is more than 2.1 mm. You are called in to help with a research study designed to assess the wear of a set of 1000 of the manufacturer's tires that have just reached the 50,000-mile mark. Use the random-number table to answer the following questions.

94163	81961	18731	89627	42895
00981	83906	68499	16409	92391
77880	41991	73241	65897	40517
27740	35486	56466	93298	71440

(a) Describe how you would use the random-number table to sample 100 of the 1000 available tires for wear.
(b) Assume you have a similar random-digit table consisting of different sets of numbers. If you selected another random sample of 100 tires and compared it to the first random sample of 100, what could you conclude with regard to which sample is best to use in the experiment? Explain.
(c) The company vice-president asks you why you have to use the random number table. He wonders why you cannot simply take the first 100 tires that became available. What should you tell him?
(d) Generate your own method of sampling tires to determine wear and state why it would be an effective design.

Part (a) would be an example of an item primarily requiring memory abilities. Parts (b)–(d) would be examples of items primarily requiring analytical, creative, and practical abilities, respectively.

3.1.2. Item development

From the summer of 2000 until the spring of 2002, item development proceeded in eight stages: (1) item development; (2) internal review; (3) review by the first expert panel of school teachers and college faculty; (4) review by consultants at the Educational Testing Service (ETS); (5) piloting of selected items by consulting teachers on the project; (6) review by the second expert panel of college faculty; (7) final review and item selection, and preparation of final assessment forms; and (8) post hoc evaluation. In March 2002, all items that had successfully passed through peer review were assembled into the assessments. Next, we describe the contents and structure of each of the augmented exams.

3.1.2.1. Psychology. For the actual AP Psychology test in 2002, students were given 70 min to complete the multiple-choice section (100 items) and 50 min to complete the open-response section (2 items) of the exam. The items were designed to cover 14 content subdomains. Table 1 gives a breakdown of each subdomain covered by the AP Psychology exam, along with the percentage of items on the test devoted to each topic. It is important to note that any more items were developed than successfully passed through the item review process. Thus, the items that did pass through the item review process were not necessarily perfectly reflective of a balanced distribution across content areas, but were the best items developed at the time of administration.[3]

The exams were scored on a scale from 1 (lowest) to 5 (highest). Consistent with the scoring of the actual AP exam, the multiple-choice section of the augmented exam was

[3] In an ideal world, we would have revised items or developed new items until a perfect balance was achieved. In reality, however, the amount of time required for the item review process is substantial, and at the end of the day, there are deadlines and issues of timing with regard to when the tests must be administered. Thus, the items chosen for the exam were the best of all items at the time. We could have deleted items in categories that were overrepresented to achieve a balance at the level of the lowest common denominator, but that would have resulted in a loss of information about otherwise strong items that successfully passed through the rigorous item review process.

Table 1
Content areas covered by the actual AP Psychology exam and the augmented AP Psychology exam and the percentage of items devoted to each

Content	MC actual AP exam[a] (%)	Total augmented AP exam (%)
Abnormal	7–9	7
Biological basis of behavior	8–10	2
Cognition	8–10	13
Developmental	7–9	1
History	2–4	4
Learning	7–9	8
Methods	6–8	14
Motivation and emotion	7–9	7
Personality	6–8	7
Sensation and perception	7–9	5
Social	7–9	9
States of consciousness	2–4	4
Testing and individual differences	5–7	8
Treatment	5–7	11

[a] *Note.* Data downloaded on 4/20/03 from http://www.collegeboard.com/ap/students/psych/cours_2002.html.

worth 50% of the total exam grade, and the open-response section was worth 50% of the total exam grade.

The augmented AP Psychology exam was designed to mimic, in many ways, the actual AP Psychology exam. A total of 85 new items were developed and piloted for the augmented AP Psychology exam, many of which were designed as open-response items. To maximize the number of items we could pilot test and to maintain some basis for equating the scores, the open-response items were distributed across two different forms, whereas the same 50 multiple-choice items appeared on both versions of the augmented AP Psychology exam. For the open-response section of the augmented exam, each version had 20 open-response items. For equating purposes, a subset of items appeared on both forms of the exam. Items 15–20 on Form A corresponded to Items 1–5 on Form B (see Appendix A for a breakdown). In the augmented exam, students were given 40 min to complete the multiple-choice section (50 items) and 110 min to complete the open-response section of the exam (20 items). Thus, the augmented exam required slightly more time than the regular exam. The difference in amount of time derived from the estimated time required to respond to our items, based on the piloting of these items with a group of students. In order for the exam to be practical, we did not want to go much over the normally allocated time, and did not.

Consistent with the actual AP Psychology exam, the new items were distributed across 14 content subdomains. At the low end of the spectrum, items assessing the Biological Basis of Behavior constituted 2% of the test items. At the high end, items from the domain of Research Methods constituted 14% of the items (see Table 1). The distribution of the items corresponded closely to those on the actual AP Psychology exam. The 85 newly developed items were distributed in the following way across the four areas of primary cognitive demand: 28% memory, 31% analytical, 19% creative,

and 22% practical. (For further information, see Stemler, Grigorenko, Jarvin, Macomber, & Sternberg, 2003a).[4]

3.1.2.2. Statistics. For the actual AP Statistics exam in 2002, students were given 90 min to complete the multiple-choice section (40 items) and 90 min to complete the open-response section of the exam (6 items). The items were designed to cover the following five content areas: (i) Experimental Design; (ii) Exploratory Data Analysis; (iii) Randomness and Sampling; (iv) Regression; and (v) Significance Testing. The augmented AP Statistics exam was designed to mimic the existing AP Statistics exam. Students were given 75 min to complete the multiple-choice section (50 items) and 75 min to complete the open-response section of the exam (6 items). Thus, the augmented exam required slightly less time than the regular exam. The difference in amount of time derived from the estimated time required to respond to our items, based on the piloting of these items with a group of students.

A total of 80 new items were developed and piloted on the augmented AP Statistics exam. Because an important condition of the study was that the teachers could use the augmented exam as a practice test for the existing exam, it was important that the amount of time required for testing be as close as possible to the existing exam. At the same time, an important goal of the project was to evaluate the psychometric properties of the newly developed items. For the augmented AP Statistics exam, many of these items were open-ended response items. Thus, to pilot-test all of the newly developed items, keeping time constant, items were distributed across three test forms. The same 50 multiple-choice items appeared on all versions of the augmented AP Statistics exam; however, to provide an easy basis for linking the scores from the various forms, and because more of the items that required pilot testing were of the open-response variety. Appendix B provides a breakdown of the number of items on each form as well as the overlap and the number of people taking each form.

Consistent with the actual AP Statistics exam, the items developed for the augmented AP Statistics exam were distributed across five content subdomains.[5] Items from the subdomain of Regression constituted a low of 8% of the items. Items from the subdomain of Exploratory Data Analysis, as well as items from the subdomain of Randomness and Sampling, constituted a high of 26% of the items each. The 80 newly developed items were distributed in the following way across the four areas of cognitive demand: 11% memory, 35% analytical, 20% creative, and 34% practical. (For further information, see Stemler, Grigorenko, Jarvin, Macomber, & Sternberg, 2003b)[4].

[4] Many more items were developed that did not successfully pass through the item-review process. Thus, the items that did pass through the item review process were not necessarily perfectly reflective of a balanced distribution, but were the best items developed at the time of administration. The matter here is a practical one. In an ideal world, we would have revised items or developed new items until a perfect balance was achieved. In reality, however, the amount of time required for the item-review process is substantial, and at the end of the day, there are deadlines and issues of timing with regard to when the tests must be administered. Thus, the items chosen for the exam were the best of all items at the time. We could have deleted items in categories that were overrepresented to achieve a balance at the level of the lowest common denominator, but that would have resulted in a loss of information about otherwise strong items that successfully passed through the rigorous item review process. We believe that our procedure maximized the amount of information obtained.

[5] The College Board did not report the percentage of items measuring each subdomain on the 2002 AP Statistics examination.

3.2. Sample

In the fall of 2000, a first wave of recruitment letters was sent out. They went to AP Psychology and Statistics teachers who had been recommended by the AP Psychology and Statistics Development Committees. This strategy failed to yield a sufficient numbers of replies. Therefore, a second wave of letters was sent out to all practicing AP teachers whose e-mail and regular mail addresses were available on different functional lists through the College Board. This mailing, carried out in the early months of 2001, included a cover letter from the College Board inviting teachers to participate in the project. The information was distributed to a large group of teachers ($N = 224$), which resulted in much greater success for our recruitment efforts.

The recruitment package comprised a cover letter from the College Board Executive Director of Advanced Placement Program and a brief outline of the program. Potential participants were informed that the purpose of the project was to promote effective teaching techniques and to help classroom teachers create formative assessments that tap higher order thinking skills. They were also told that participation in the project would require them to arrange for the administration of the newly developed practice exam (the enhanced AP exam) sometime in the spring of the following academic year (spring 2002), prior to the actual AP exam. All teachers were compensated monetarily for their participation at the rates specified by ETS and the College Board in their collaboration with AP teachers.

In response to our recruitment efforts, a total of 33 AP Psychology and 23 AP Statistics teachers volunteered to participate in this study. Participating teachers were drawn from across the U.S.A. and represented a total of 19 different States, with each participating teacher representing a different school. The number of years of prior teaching experience for AP Psychology teachers ranged from 1 to 10 (mean = 7.08, mode = 10), with 20 teachers having experience as readers for the AP national test. Participating AP Statistics teachers ranged in years of teaching experience from 0 to 6 years (mean = 3.5, mode = 5), with nine having experience as readers for the AP national test.

Table 2 presents the demographic breakdown of the students taking the augmented exams in AP Statistics and AP Psychology. Note that, in this project, no students took both the Statistics and Psychology exams. We were able to obtain information on the eth-

Table 2
Demographic breakdown for participating students

	Statistics	Psychology
Gender		
Male	232	349
Female	284	635
Missing	117	278
Ethnicity		
White	210	338
Black	13	14
Hispanic	11	23
Asian	54	60
Other	18	17
Missing	327	810
Total *N*	633	1262

nicity of 452 of the 1262 students (36%) taking the augmented AP Psychology exam. The ethnic breakdown of these 452 students was 75% White, 3% Black, 5% Latino, 13% Asian, and 4% Other. These numbers are roughly comparable to the proportion of students from each ethnic group who took the actual AP Psychology exam in 2002 (cf. 70% White, 5% Black, 7% Latino, 12% Asian, and 4% Other). Despite the low response rate, an analysis of the test score data revealed no statistically significant differences in their overall score on the augmented AP Psychology exam between those reporting ethnic information and those not reporting ethnic information.

Of the 633 students who took the augmented AP Statistics exam, we were able to obtain ethnic background information for 306 students (48%). The ethnic breakdown of these 306 students was 68% White, 4% Black, 4% Latino, 18% Asian, and 6% Other. These numbers are also nearly identical to the proportion of students from each ethnic group who took the actual AP Statistics exam in 2002 (cf. 68% White, 4% Black, 6% Latino, 17% Asian, and 4% Other). An analysis of the data revealed that those for whom we did not have ethnic information showed significantly lower levels of achievement overall on the augmented AP Statistics exam than either White students or Asian students, but did not show statistically significant differences in achievement from Black or Latino students.

4. Results

4.1. Main findings for research question 1: Psychometric properties of the instruments

The data were analyzed using both classical test theory (Crocker & Algina, 1986) and Rasch measurement (Bond & Fox, 2001; Rasch, 1960/1980; Wright & Stone, 1979). In this article, we report the results from the Rasch analysis as they provide the most precise estimates of student ability (Bond & Fox, 2001; Smith, 1996, 2001; Wright & Stone, 1979). Evidence in the psychometric literature suggests that Rasch estimates are more precise estimates of latent abilities than classical estimates (Smith, 1996, 2001). Classical analyses assume that the true scores of all test takers are measured equally well (i.e., they are assigned a uniform standard error of estimate). With Rasch analyses, each participant may be assigned a distinct standard error of estimate depending on which items the participant got correct or incorrect. This feature greatly increases the precision of estimates of latent abilities (or true scores) as compared with classical test theory methods. (For further details see Smith, 2001; Smith, 1996.)

The data for each of the exams were analyzed using the many-facets Rasch model (Linacre, 1988, 1994; Linacre, Wright, & Lunz, 1990). This approach has several advantages over classical test theory. First, it puts each of the items onto a linear scale. Second, it effectively deals with incorporating information from multiple raters, correcting for rater severity in the ability estimate. Third, it is an effective technique for combining the results of multiple-choice and open-response items into a single ability estimate.

A series of five separate Rasch analyses were run for each domain (i.e., Psychology and Statistics). The first analysis included all items for an overall ability estimate. The next four analyses were designed to generate ability estimates for each subscale using only items that were explicitly designed for to measure the said process (e.g., scale 1 = memory items, scale 2 = analytical items).

There is always a fundamental tension in test construction between the desire for the measurement of a unidimensional construct (e.g., Statistics ability), and the recognition

that the construct itself may be divided into various subdomains on the basis of content areas (e.g., probability, sampling) or kinds of processes (e.g., creative, analytical). Thus, few would ever argue that we can attain a purely unidimensional construct. Nevertheless, Rasch analysis does not require this strict interpretation to be useful. In fact, the underlying dimension driving the ability estimate is Statistics ability. This overarching construct comprises subcomponents—domain-specific knowledge as well as specific cognitive skills. They are useful both for the purposes of test construction, as well as for diagnostic purposes (such as the identification of areas in which participants exhibit strengths and weaknesses).

Consistent with the way the actual AP exam is scored, the multiple-choice and open-response sections of the augmented exams received equal weighting. Each student received an ability estimate for the overall exam, as well as an ability estimate for each of the subscales assessing memory, analytical, creative, and practical skills. To aid interpretation of the scale and avoid the use of the negative logit values associated with the Rasch approach, each of the Rasch ability estimates was rescaled to have a mean of 250 and a *SD* of 50. In addition, five proficiency levels were developed for each scale, corresponding to students at quintile intervals (e.g., students scoring 2 on the scale represent students scoring from the 21st to the 40th percentile).

4.1.1. Content-related validity evidence

Content-related validity evidence for the augmented exams was gathered following the eight-step process described above and detailed elsewhere (Stemler et al., 2003a, 2003b). To summarize briefly, items were systematically developed based on Sternberg's theory of successful intelligence. They were then sent out to expert teachers and test developers for review. The evaluation criteria were (a) the extent to which each item accurately tapped into memory, analytical, creative, or practical abilities and (b) its central position in the content domain. Moreover, the content of the items was thoroughly evaluated.

Once a set of multiple-choice and open-ended items had been reviewed and revised internally, two separate test forms were created and sent out for evaluation and review to a panel of six active AP teachers and six college faculty in statistics ($N = 12$) and to eight active AP teachers and eight college faculty in psychology ($N = 16$). The results of the evaluation indicated that the majority of the items fitted the described item specifications. All relevant comments from the reviewers were incorporated in the new revision of the items.

After the completion of the first external review, the items were revised once again and sent out for another round of reviews, this time to an AP psychology coordinator and an AP statistics coordinator at Educational Testing Service (ETS). The reviewers were asked to evaluate the items and provide detailed comments, which were incorporated in yet another revision of the items.

Next, a number of AP teachers consulting on the project ($N = 16$) were asked to pilot the items with their students and to provide their feedback and the students' comments on the items. All relevant suggestions were incorporated.

Finally, four independent consultants, individuals who worked at universities and who had expertise in the relevant content area as well as expertise in item development, were hired to review the items for content, clarity, and potential bias. The reviewers' comments were then used to modify items before creating the final exam booklets.

4.2. Criterion-related validity evidence

Criterion-related validity evidence was gathered by correlating scores from the augmented exam with scores from the actual AP exam for a subset of individuals. The results, reported in Table 3, were generally in line with our research hypothesis. The analytical subscale was correlated most highly with the actual AP exam score ($r = .54$, $p < .01$), whereas the practical subscale was least correlated with the actual AP exam score ($r = .33$, $p < .01$). A test for the difference between two dependent correlation coefficients from the same sample (Blalock, 1972, p. 407) was conducted for each pair of correlation coefficients (e.g., memory vs. practical). A total of six paired comparisons were conducted. The results showed that the difference between the correlation of the memory subscale with the actual exam score and the correlation of the analytical subscale and the actual exam score was nonsignificant. The differences between all other pairs of correlations were statistically significant at the .05 α level. A Bonferroni adjustment for 6 multiple-comparisons yielded a critical t value of 2.64 for a two-tailed test at the .05 α level. The differences in correlations provide some evidence that the subscales are measuring constructs that differ significantly from each other (memory–analytical, N.S.; memory–creative $t = 4.04$, $p < .01$; memory–practical $t = 6.83$, $p < .01$; analytical–creative $t = 5.77$, $p < .01$; analytical–practical $t = 7.85$, $p < .01$; creative–practical $t = 2.53$, $p < .05$).

In addition, the results in Table 4 reveal that the total score from the augmented version of the AP Statistics exam and the existing AP Statistics exam were moderately correlated ($r = .49$, $p < .001$). We also hypothesized that the correlations involving the memory subscale, the analytical subscale, and the actual AP Statistics exam would be

Table 3
Bivariate correlations between the actual AP score and the overall and subscale scores of the augmented AP Psychology exam

Scale	1	3	4	5	6
1. Actual AP exam	1.00				
2. Augmented AP exam	0.61**				
3. Memory subscale	0.52**	1.00			
4. Analytic subscale	0.54**	0.47**	1.00		
5. Creative subscale	0.41**	0.40**	0.53**	1.00	
6. Practical subscale	0.33**	0.42**	0.43**	0.41**	1.00

N for actual AP exam with other scales = 733; N for other scales = 1262.
** Correlation is significant at the 0.01 level (two tailed).

Table 4
Bivariate correlations between the actual AP score and the overall and subscale scores of the augmented AP Statistics exam

Scale	1	3	4	5	6
1. Actual AP exam	1.00				
2. Augmented AP exam	0.49**				
3. Memory subscale	0.45**	1.00			
4. Analytic subscale	0.39**	0.35**	1.00		
5. Creative subscale	0.36**	0.43**	0.43**	1.00	
6. Practical subscale	0.43**	0.44**	0.46**	0.55**	1.00

N for actual AP exam with other scales = 393; N for subscales = 633.
** Correlation is significant at the 0.01 level (two tailed).

greater than the correlations involving the creative subscale, the practical subscale, and the actual AP exam, as the existing exam consists mostly of items tapping memory and analytical skills. The results presented in Table 4 were generally in line with our predictions. The memory subscale was correlated most highly with the actual AP exam score ($r = .45$, $p < .01$), whereas the creative subscale was least correlated with the actual AP exam score ($r = .36$, $p < .01$). However, after correcting for multiple comparisons, no significant differences between coefficients were found.

Because multiple comparisons were made for each test, the Bonferroni correction procedure was used. Although the Bonferroni correction diminishes the probability of Type I error, it will inflate the probability of Type II error. Given that this is the first study of its kind, the decision to use this correction may have been overly conservative.

4.2.1. Construct-related validity evidence

Evidence of the construct validity of the exams was provided through various statistics from the many-facets Rasch model. The Rasch person-reliability and item-reliability estimates for the augmented AP Psychology exam are presented in Table 5. The Rasch reliability estimates are interpreted in the same way as Cronbach's α, but provide estimates that are more precise (see Fisher, 1992; Smith, 2001). Table 6 presents the same information for the augmented AP Statistics exam.

In general, separation values greater than 2.0 indicate that the scale is working as desired (Bond & Fox, 2001). Larger item-separation values indicate that the items are targeted at varying levels of difficulty rather than all being targeted at approximately the same level of difficulty. Within the context of Rasch measurement, high item-reliability values suggest that if the same test were given to a different sample of participants with similar abilities, the relative positioning of item difficulties would be expected to remain constant.

Table 5 reveals that the item separation index for the overall scale of the augmented AP Psychology exam was 18.96, indicating that there was a substantial spread of items along the logit scale. In other words, the items were not all targeted around a limited set of ability levels, or at the same level of difficulty. The item reliability for the overall scale was 1.0. The average person ability was 0.06 logits, meaning that the ability level of this sample of test takers was almost perfectly matched to the difficulty level of the items on the test.

Fig. 1 presents an item map for the overall augmented AP Psychology exam and Fig. 2 presents an item map for the overall augmented AP Statistics exam. It is easy to see from the item map that there is a spread of items associated with each of the process areas. The items on the item map are labeled with a prefix indicating their specific process area (e.g., M14 means that item 14 assesses a memory process).

The 12 raters who scored the open-response items on the augmented AP Psychology exam differed somewhat in terms of their severity and the range in severity differed across subscales. Within the domain of psychology, the range in rater severity was approximately 1.75 logits for the memory subscale, 1.25 logits for the analytical subscale, and 2.3 logits for the creative and practical subscales.

In addition, the rater fit statistics show that raters were fairly consistent in their application of the scoring rubric, with only 2 raters exhibiting greater than expected variation in their ratings on each of the memory, analytical, and creative subscales. Interestingly,

Table 5
Person and item estimates: augmented AP Psychology exam

	Overall	Memory	Analytical	Creative	Practical
Psychology—item summary					
Summary of item estimates					
Mean	0.00	0.00	0.00	0.00	0.00
SD	1.23	1.47	1.02	1.14	1.09
Reliability of estimate	1.00	1.00	1.00	1.00	1.00
Separation	18.96	18.42	17.04	18.74	18.19
N of items	85	24	26	16	19
Summary of fit statistics					
Infit mean square					
Mean	1.20	1.10	1.10	1.10	1.00
SD	0.40	0.30	0.20	0.30	0.10
Outfit mean square					
Mean	1.20	1.20	1.10	1.20	1.10
SD	0.60	0.30	0.30	0.40	0.20
Infit z					
Mean	2.10	2.60	1.50	1.30	1.40
SD	3.40	3.60	3.70	4.70	4.20
Outfit z					
Mean	2.20	2.70	1.60	1.90	1.50
SD	3.50	3.30	3.70	4.90	4.10
Psychology—person summary					
Summary of person estimates					
Mean	0.06	−0.22	0.07	0.31	0.11
SD	0.66	0.99	0.85	1.07	0.94
Reliability of estimate	0.92	0.79	0.82	0.75	0.76
Separation	3.28	1.94	2.17	1.75	1.79
N of test-takers	1262	1262	1262	1262	1262
Summary of fit statistics					
Infit mean square					
Mean	1.30	1.30	1.10	1.00	1.00
SD	0.50	0.80	0.60	0.70	0.60
Outfit mean square					
Mean	1.20	1.20	1.10	1.10	1.10
SD	0.40	0.80	0.50	1.00	0.80
Infit z					
Mean	1.40	0.40	0.00	−0.30	−0.20
SD	2.00	1.60	1.60	1.50	1.50
Outfit z					
Mean	0.80	0.10	0.00	−0.10	−0.10
SD	1.60	1.10	1.30	1.30	1.30

although the rater severity spans the greatest range for the practical items, the rater fit statistics reveal that all of the raters remained faithful to their interpretation of the scoring rubric (i.e., there were no misfitting raters).

Table 6
Person and item estimates: augmented AP Psychology exam

	Overall	Memory	Analytical	Creative	Practical
Statistics—item summary					
Summary of item estimates					
Mean	0.00	0.00	0.00	0.00	0.00
SD	1.05	1.33	1.17	1.02	0.95
Reliability of estimate	0.99	0.99	1.00	0.99	0.99
Separation	12.87	12.79	14.22	11.70	11.27
N of items	80	9	27	15	26
Summary of fit statistics					
Infit mean square					
Mean	1.10	1.00	1.20	1.10	1.00
SD	0.50	0.30	0.20	0.20	0.10
Outfit mean square					
Mean	1.20	0.30	1.30	1.10	1.10
SD	0.60	4.70	0.30	0.30	0.20
Infit z					
Mean	1.20	1.20	3.60	1.20	0.80
SD	3.00	0.50	2.90	4.10	3.50
Outfit z					
Mean	1.40	0.80	3.80	1.10	1.30
SD	3.10	5.00	3.00	4.40	3.80
Statistics—person summary					
Summary of person estimates					
Mean	−0.23	0.54	−0.36	0.08	0.02
SD	0.67	1.31	0.79	1.14	0.98
Reliability of estimate	0.94	0.57	0.91	0.82	0.85
Separation	3.90	1.15	3.18	2.11	2.38
N of test-takers	633	633	633	633	633
Summary of fit statistics					
Infit mean square					
Mean	1.70	0.90	1.50	0.90	1.00
SD	0.90	0.50	0.80	0.60	0.40
Outfit mean square					
Mean	1.40	1.20	1.30	1.10	1.00
SD	0.60	1.30	0.70	0.70	0.50
Infit z					
Mean	2.50	−0.40	1.30	−0.50	−0.30
SD	3.10	1.00	2.30	1.50	1.60
Outfit z					
Mean	1.80	−0.10	0.80	−0.20	−0.10
SD	2.60	1.00	1.70	1.30	1.40

For statistics, the six raters who scored the various open-response items on the augmented AP Statistics exam differed somewhat in terms of their severity and the range in severity differed across subscales. Within the domain of statistics, the range in rater severity was approximately 0.37 logits for the memory subscale, 1.64 logits for the

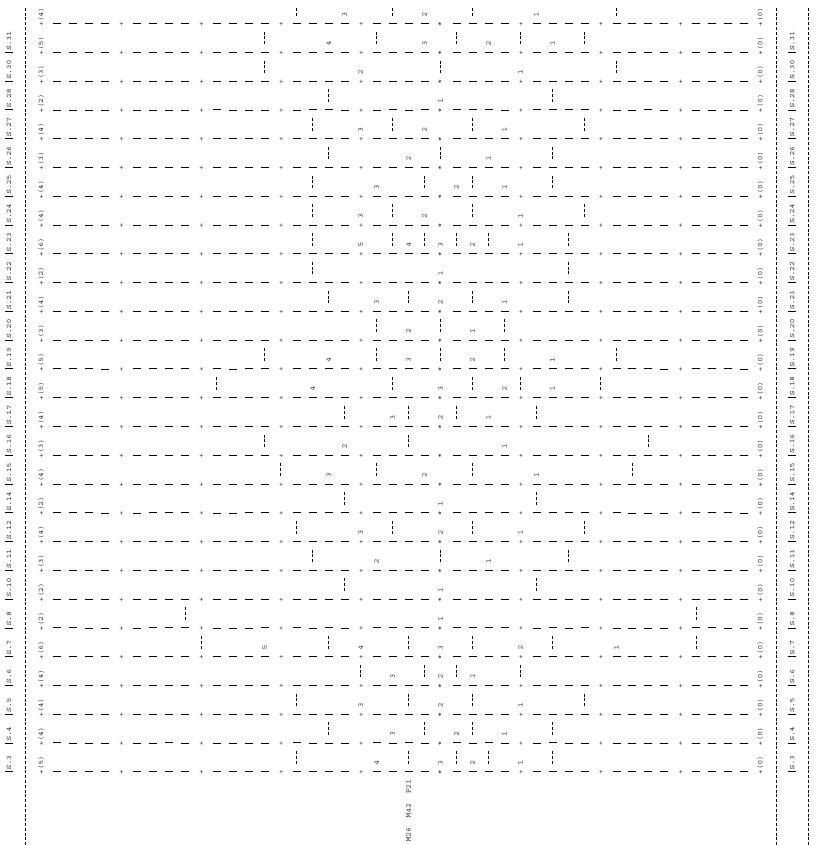

Fig. 1. Item map for the augmented AP Psychology exam.

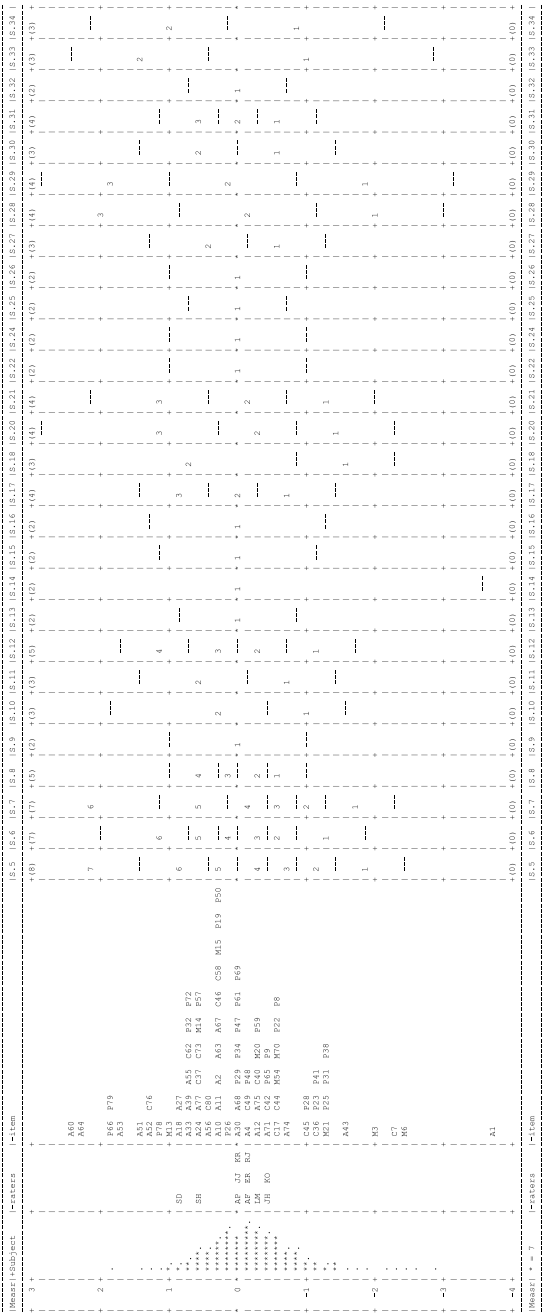

Fig. 2. Item map for the augmented AP Statistics exam.

analytical subscale, 0.86 logits for the creative subscale, and 1.63 logits for the practical subscale.

In addition, the rater fit statistics show that raters were fairly consistent in their application of the scoring rubric, with only 2 raters exhibiting greater than expected variation in their ratings on the analytical scale, and no raters exhibiting misfit on the each of the memory, creative, or practical subscales.

4.2.2. Summary

Overall, the various sources of evidence support the validity of using the theory of successful intelligence as a basis for creating augmented exams in AP Psychology and Statistics. Items tapping memory, analytical, creative, and practical skills are distinguishable on the basis of their content. In addition, the subscales containing items tapping each different cognitive process showed the expected pattern of correlations with the actual AP exam results. Finally, the various subscales exhibited acceptable levels of internal consistency and other item statistics, with the single exception of the memory subscale of the Statistics exam, whose internal consistency (for persons, not items) was substandard.

4.3. Main findings for research question 2: Individual differences in cognitive processes

To examine the extent to which individuals exhibited different profiles of strengths and weaknesses across the four cognitive-skill areas under investigation, we conducted a cluster analysis using the Q-factor analysis approach (Hair, Anderson, Tatham, & Black, 1998). Under this approach, students with similar patterns of relative strengths and weaknesses, regardless of differences in absolute levels of achievement, are identified as belonging to the same cluster. The results will now be discussed for each content area.

4.3.1. Psychology

Using principal components extraction with a promax rotation, we obtained three factors with eigen values greater than 1.0 that accounted for 100% of the variance in the dataset. The three factors correspond to three distinct profiles of achievement primarily found in the dataset. Each participant thus had a factor loading corresponding to each of the three extracted factors. Table 7 presents an abridged structure matrix of factor loadings for each participant in the augmented AP Psychology exam.

Table 7 illustrates that some students had a clear, high positive loading on a single factor, whereas other students had factor loadings that were high, but in the negative direction.

To gain a deeper appreciation for the meaning behind each of these factors, Fig. 3 presents the profiles of achievement for 12 participants. The four participants in the first row had high positive loadings on the first factor; the four participants shown in the second row had high positive loadings on the second factor; and the four participants shown in the third row had high positive loadings on the third factor.

An examination of Fig. 3 reveals that those participants with high loading on the first factor tend to exhibit profiles of achievement with relatively low scores on the memory subscale compared with their achievement on the analytical, creative, and practical subscales. At the same time, their scores on the analytical, creative, and practical subscales were roughly equivalent.

Table 7
Abridged output of factor loadings for each participant: augmented AP Psychology exam

SID	Component		
	1	2	3
K_01060061	**0.998**		
K_00020051	**0.997**		
K_00080081	**0.996**		
K_02080015	**0.996**		
K_02020017	0.996		
K_00120027	0.995		
…			
K_02030014		−0.982	
K_02100027		−0.982	
K_02080064		**0.977**	
K_00120034		**0.976**	
K_00080003		**0.976**	
K_00080102		**0.976**	
K_00140006		−0.975	
…			
K_02080019			−0.996
K_00140017			**0.995**
K_00010021			**0.995**
K_02090013			**0.995**
K_00020083			**0.993**
K_00070008			0.992
K_02100017			0.991
…			

SIDs listed in bold correspond to person profiles show in Fig. 3.

Students with high loadings on the second factor (second row of Fig. 3) tended to have weaker levels of achievement on items tapping practical thinking skills; however, their achievement on the memory, analytical, and creative scales was roughly equivalent to one another.

Finally, students with high loadings on the third factor (third row of Fig. 3) showed a pattern of relative strength on creative items, relative weakness on the analytical items, and moderate achievement on the memory and practical items.

Although three principal components were extracted from the Q-factor analysis, it is important to note that some participants had high positive loadings on a factor, and other participants had high negative loadings on the same factor. Participants with high negative loadings on the first factor show a pattern of achievement that is the mirror image of those students with positive loadings on that factor. Specifically, participants with high negative loadings on the first factor exhibit relatively strong achievement on memory items, and lower but relatively equal achievement on analytical, creative, and practical items. Thus, although three factors were extracted, these factors yielded six distinct profiles of achievement.

Table 8 presents a summary of the number of participants whose profile is associated with each of the six empirically distinguishable profiles of achievement. For example, 30% of the 1262 participants exhibited a profile of achievement associated with a high positive loading on Factor 1 (relative weakness on memory skills), whereas 19% of participants exhibited a profile of achievement associated with high positive loadings on Factor 3

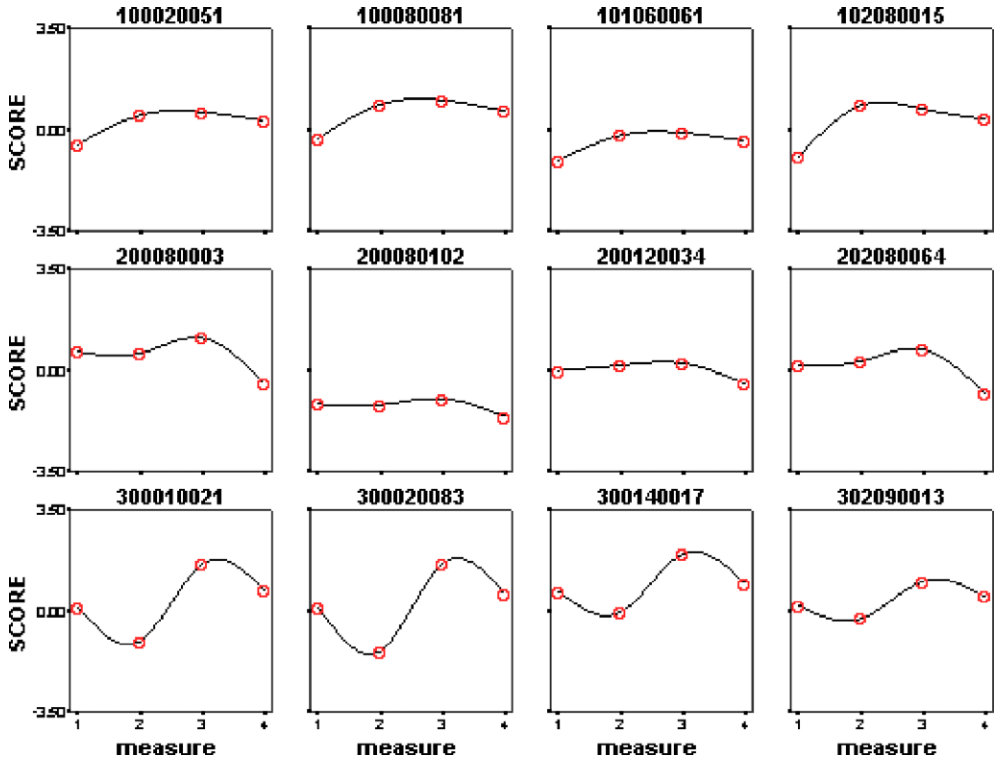

Fig. 3. Augmented AP Psychology exam—exemplary empirical profiles of achievement. 1 = Memory subscale score (logits), 2 = analytical subscale score, 3 = creative subscale score, 4 = practical subscale score.

Table 8
Summary of participants associated with each cluster analysis profile: augmented AP Psychology exam

	Psychology			
	Positive loading		Negative loading	
	N	%	*N*	%
Cluster 1	377	30	122	10
Cluster 2	211	17	201	16
Cluster 3	243	19	108	9

(i.e., relative strength in creative skills and relative weakness in analytical skills). These findings suggest that tests developed with items measuring primarily memory and analytical skills would fail to detect the relative strengths of many participants.

4.3.2. Statistics

Using principal components extraction with a promax rotation, we obtained three factors with eigenvalues greater than 1.0 that accounted for 100% of the variance in the dataset. The three factors correspond to three distinct profiles of achievement primarily found in the dataset. Each participant thus had a factor loading corresponding to each of the three extracted factors. Table 9 presents an abridged structure matrix of factor loadings for each participant in the augmented AP Statistics exam.

Table 9
Abridged output of factor loadings for each participant: augmented AP Statistics exam

SID	Component		
	1	2	3
K_05020027	**0.998**		
K_03040019	**0.998**		
K_03090002	**0.997**		
K_05030039	**0.997**		
K_03040037	−0.995		
K_03010012	0.995		
…			
K_04070009		**0.999**	
K_03010017		**0.997**	
K_05090001		**0.995**	
K_04060021		**0.995**	
K_03070012		−0.991	
K_03050010		0.990	
…			
K_05060016			**0.999**
K_04080004			**0.996**
K_05090016		0.318	**0.993**
K_05020042	0.333	0.314	**0.990**
K_03070004		0.361	0.989
K_03090004			0.989
K_05020005			0.984
K_03040063			0.982
…			

SIDs listed in bold correspond to person profiles show in Fig. 4.

Table 9 illustrates that some students had a clear high positive loading on a single factor, whereas other students had factor loadings that were high, but in the negative direction. The findings from the Q-factor analysis show that there are three empirically distinguishable profiles of achievement that come out in the data.

For illustration purposes, Fig. 4 presents the profiles of achievement for 12 participants. The four participants in the first row had high positive loadings on the first factor. The four participants show in the second row had high positive loadings on the second factor, and the four participants show in the third row had high positive loadings on the third factor.

An examination of Fig. 4 reveals that those participants with high loading on the first factor tend to exhibit relatively high scores on the memory subscale compared with the analytical, creative, and practical subscales. At the same time, the relative achievement on the analytical, creative, and practical subscales was roughly equivalent.

Students with high loadings on the second factor (second row of Fig. 4) tended to have weaker levels of achievement on items tapping analytical thinking skills and relatively strong achievement on items assessing creative skills.

Finally, students with high loadings on the third factor (third row of Fig. 4) showed a pattern of relative weakness on items assessing analytical thinking skills. Yet, their average achievement on items tapping memory, creative, and practical skills was roughly equivalent.

Although three principal components were extracted from the Q-factor analysis, it is important to note that some participants had high positive loadings on a factor, and other

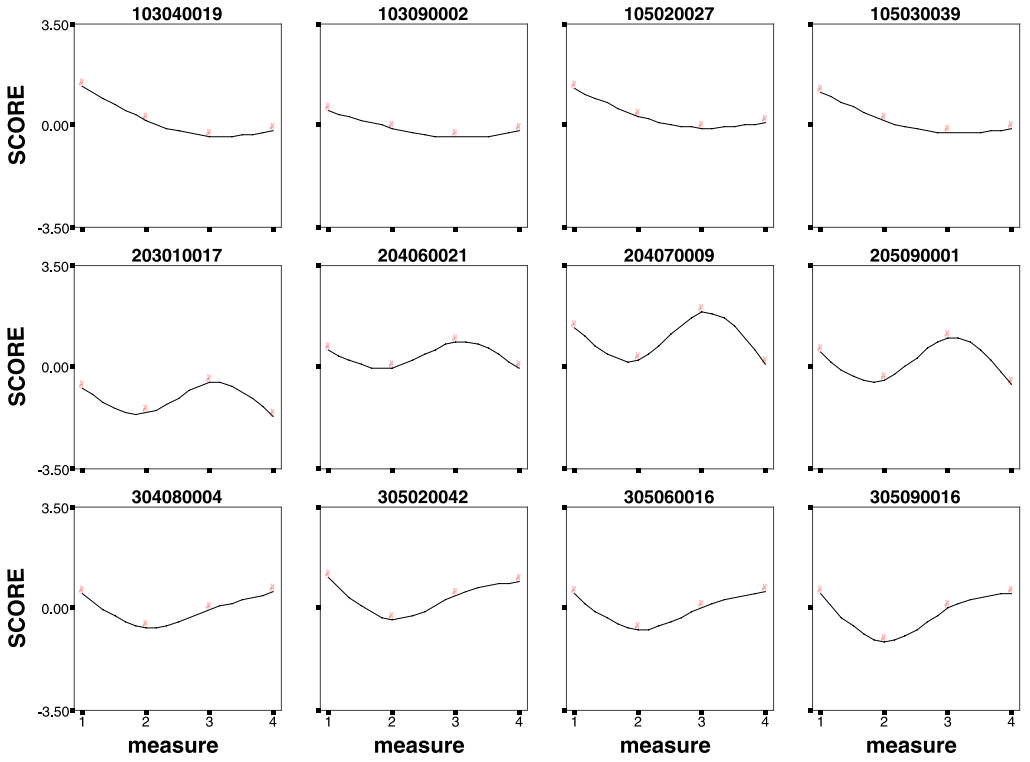

Fig. 4. Augmented AP Statistics exam—exemplary empirical profiles of achievement. 1 = Memory subscale score (logits), 2 = analytical subscale score, 3 = creative subscale score, 4 = practical subscale score.

participants had high negative loadings on the same factor. Participants with high negative loadings on the first factor show a pattern of achievement that is the mirror image of those students with positive loadings on that factor. Specifically, participants with high negative loadings on the first factor exhibit relatively weak achievement on memory items, and lower but relatively equal achievement on analytical, creative, and practical items. Thus, although three factors were extracted, these factors yielded six distinct profiles of achievement.

Table 10 presents a summary of the number of participants whose profile is associated with each of the six empirically distinguishable profiles of achievement. For example, 24% of the 633 participants taking the augmented AP Statistics exam exhibited a profile of achievement associated with a high positive loading on Factor 3 (relative

Table 10
Summary of participants associated with each cluster analysis profile: augmented AP Statistics exam

	Statistics			
	Positive		Negative	
	N	%	*N*	%
Cluster 1	184	29	76	12
Cluster 2	142	22	43	7
Cluster 3	155	24	33	5

weakness on analytical skills), whereas 22% of participants exhibited a profile of achievement associated with high positive loadings on Factor 2 (i.e., relative strength in creative skills). These findings suggest that tests developed with items measuring primarily memory and analytical skills would fail to detect the relative strengths of many participants.

4.3.3. Summary

Overall, the findings from the cluster analyses presented here provide empirical support for the assertion that individuals exhibit different profiles of strengths and weaknesses across memory, analytic, creative, and practical skills. These findings suggest that tests that measure a limited range of processes may fail to detect the strengths of a substantial proportion of test-takers.

4.4. Main findings for research question 3: Ethnic differences in achievement

Table 11 presents the mean scale scores and standard deviation by student ethnicity on the actual AP exam, as well as for the overall augmented AP Psychology exam and each of the cognitive processing subscales. Recall that the actual AP exam is scored on a scale of 1 (lowest) to 5 (highest); the mean for all test takers on the augmented exam and each of the subscales is 250, with a *SD* of 50 points.

A psychometric analysis of item difficulty found that items across all subscales of the augmented AP exams represented a broad range of difficulty levels. Consequently, higher ability estimates on creative and practical subscales are independent of the difficulty level of the items. In other words, the results demonstrated that creative and practical items are just as difficult (from a psychometric perspective) as items tapping analytical and memory items across the entire sample of test takers. This finding held for items on both the augmented AP Psychology exam and the augmented AP Statistics exam.

4.4.1. Psychology

The results from Table 11 indicate that, on the actual AP Psychology exam administered in 2002, the standardized difference in achievement between Black students and

Table 11
Augmented AP Psychology exam subscale scores by ethnicity

Scale	White		Black		Latino	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Actual AP exam (2002)	3.43	(1.23)	2.52	(1.29)	2.70	(1.31)
N	35,386		2724		3324	
Augmented exam	250	(49)	238	(56)	233	(44)
Memory subscale	252	(46)	254	(50)	230	(48)
Analytic subscale	247	(47)	234	(56)	238	(47)
Creative subscale	253	(47)	252	(57)	237	(54)
Practical subscale	253	(51)	231	(45)	247	(40)
N	338		14		23	

Note. The actual AP exam is scored on a 1–5 performance scale. We did not have access to students' raw scores, but rather only to the overall performance level scores. By contrast, the data from the Augmented exam are based on raw scores that were scaled using Rasch measurement. Each of the scales of the Augmented exam has a mean of 250 and a *SD* of 50 when the data for all participants are analyzed together.

Whites was fairly large, with Black students scoring nearly three-quarters of a standard deviation below White students (Cohen's $d = -0.72$). By way of comparison, White students also tended to have the highest scores overall on the augmented AP Psychology exam. Black students in our sample scored approximately one-quarter of a standard deviation (Cohen's $d = -0.23$) lower than the mean of White students on the overall exam, and on the analytical subscale ($d = -0.25$). A key finding, however, is that the effect size difference between Black students and White students was virtually non-existent for both the creative subscale ($d = -0.02$) and the memory subscale ($d = 0.04$). To our surprise, the biggest gap in achievement between this sample of Black students and White students was observed on the practical subscale ($d = -0.45$).

In addition, a comparison of the standardized difference in achievement between Latino students and White students on the actual AP exam reveals that Latino students scored a little more than half a standard deviation below White students ($d = -0.58$). By way of comparison, Latino students scored about one-third of a standard deviation below White students on the overall augmented AP Psychology scale ($d = -0.37$). The largest difference between Latino students and White students was observed on the memory subscale of the augmented AP Psychology exam, wherein Latino students scored approximately one-half a standard deviation below the White students ($d = -0.47$). Yet, the effect size difference between Latino students and White students was somewhat lower on the creative subscale ($d = -0.32$), and substantially lower on the practical subscale ($d = -0.13$).

In general, then, our exam reduced differences between ethnic groups relative to the actual AP exam. Thus, it appears simultaneously to measure a broader range of skills and to reduce differences between groups.

4.4.2. Statistics

Table 12 presents the mean scale scores and standard deviation by ethnicity for the actual AP Statistics exam from 2002, as well as the augmented AP Statistics exam and each of the cognitive processing subscales. The results indicate that Black students scored approximately three quarters of a standard deviation below White students ($d = -0.77$) on the actual AP exam. By way of comparison, the results from Table 12 show that White students and Asian students tended to have the highest scores overall on the augmented exam

Table 12
Augmented AP Statistics exam subscale scores by ethnicity

Scale	White		Black		Latino	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Actual AP exam (2002)	2.84	(1.29)	1.90	(1.14)	2.05	(1.20)
N	33,368		1950		2879	
Augmented AP exam	262	(38)	207	(60)	232	(61)
Memory subscale	257	(47)	220	(47)	244	(47)
Analytic subscale	258	(37)	200	(72)	210	(67)
Creative subscale	263	(42)	248	(57)	243	(43)
Practical subscale	263	(43)	227	(48)	240	(47)
N	210		13		11	

Note. The actual AP exam is scored on a 1–5 performance scale. We did not have access to students' raw scores, but rather only to the overall performance level scores. By contrast, the data from the Augmented exam are based on raw scores that were scaled using Rasch measurement. Each of the scales of the Augmented exam has a mean of 250 and a *SD* of 50 when the data for all participants are analyzed together.

for Statistics. Black students in the sample scored one full standard deviation below the mean of the White students on the overall subscale ($d = -1.10$). In addition, we observed a similar gap in achievement on the analytical ($d = -1.01$) subscale. The effect size difference between Black students and White students was somewhat lower on both the memory ($d = -0.79$) and practical subscales ($d = -0.77$), and was drastically reduced on the creative subscale ($d = -0.30$). To put these findings into context, the effect size difference between White students and Black students on the actual exam ($d = -0.77$) means that a White student scoring at the 50th percentile of the White student distribution would outscore 78 percent of the Black students taking the same exam. By contrast, a White student scoring at the 50th percentile on the creative subscale would outscore 63 percent of Black students on the creative subscale.

Interestingly, whereas Black students tended to exhibit smaller differences in achievement on the creative subscale, Latino students exhibited smaller differences in achievement on the practical subscales. Latino students scored approximately one standard deviation below the White students on the analytical subscale of the augmented exam ($d = -0.89$). Yet, the effect size difference between Latino students and White students was somewhat lower on the creative ($d = -0.47$) and practical subscales ($d = -0.50$), and much lower on the memory subscale ($d = -0.28$). These results can be compared with the results of the actual AP Statistics exam in 2002 Table 12 in which Latino students scored approximately two thirds of a standard deviation below White students ($d = -0.63$).

4.4.3. Summary

In summary, the augmented exam thus generally reduced differences between groups on the Statistics exam, as on the Psychology exam.

5. Discussion

This study has provided some suggestive answers to our three research questions. First, the results indicate that it is possible to create psychometrically sound instruments based on the theory of successful intelligence that measure students' cognitive skills in the context of AP Psychology and AP Statistics.

Second, students do exhibit somewhat different profiles of strengths and weaknesses across different cognitive skills, regardless of content domain. Indeed, the results shown here demonstrate that a subset of students even exhibit extreme differences in their achievement, scoring at the lowest levels on one skill and the highest levels of other skills. Thus, tests that measure only one cognitive skill may tend to miss important information about individuals with strengths in other cognitive skill areas. Tests that measure only a narrow range of cognitive skills may therefore lead to less valid inferences of student ability.

Third, the results of an analysis of ethnic differences in achievement show that ethnic minority students appear to benefit from assessments that measure a broader range of cognitive skills. Some evidence for this assertion comes from the fact that the pervasive achievement gap between Black students and White students that has been consistently observed across many achievement tests was also observed on the analytical subscales for both of the augmented exams created for this study. Yet, the Black–White test score differences were eliminated or greatly reduced on the creative subtests for both AP Psychology and Statistics. In addition, the differences in achievement between White stu-

dents and Latino students were greatly reduced on the practical subtest of the augmented exams. Although the sample size on which these findings are based is rather small and the sample is not necessarily representative, the findings suggest that tests that measure only a limited range of cognitive skills, or that make no effort to explicitly balance the range of cognitive skills assessed, may inadvertently not reveal the full range of important skills of at least some members of particular ethnic groups. To ensure equity for individuals and ethnic groups, it is important to develop tests that assess a wide range of cognitive skills.

5.1. Limitations

One important limitation of the study relates to the sample sizes of the ethnic minority students in the study. Although the results of our analyses are suggestive, they must be interpreted with caution, given the small *n*'*s* of the ethnic minority students on which the results are based. Nevertheless, the results of this study provide first estimates of the effect sizes. Thus, future researchers may use these data to calculate sampling strategies for future studies that are powerful enough to detect the desired effects.

The small number of ethnic minority students taking our exam is somewhat reflective of a larger problem with the AP program itself. Indeed, proportionately, the percentage of students from various ethnic groups taking our exams was nearly identical to the percentage of students from each ethnic group taking the actual AP exam. A great challenge facing the AP program is in recruiting minority students to participate. We also cannot and do not claim that our data are fully representative of the various groups that we tested.

5.2. Directions for future research

The results of this study suggest that the theory of successful intelligence provides a useful basis for test construction. Future studies that replicate this approach in different content domains are warranted to further test the generalizability of this approach. In addition, the findings from the examination of ethnic group differences must be replicated with a larger sample of students before any firm conclusions can be drawn. Future studies should be conducted that explicitly over-sample students from the underrepresented minority groups of interest to more fully examine the extent to which the preliminary findings presented here hold.

6. Conclusions

Overall, the results reported here are promising. Explicitly balancing tests for both content and cognitive processing skill appears to be potentially beneficial at both the individual and group levels. At the individual level, a profile-oriented approach to scoring may lead to the identification of students with strengths in areas not traditionally measured by tests of achievement. At the group level, broadening the range of cognitive skills assessed on tests of achievement may lead to greater equity and increased validity in using the results to make inferences about students' level of content mastery. Broadening the range of cognitive skills assessed may allow us to create a more comprehen-

sive assessment system, whereby diverse cognitive-processing skills are valued and rewarded. Educational institutions may be better able to select students who exhibit a range of cognitive processing skills, thereby enriching the academic experience of all students and creating greater equity within the context of a high-stakes testing program.

Acknowledgments

Preparation of this article was supported by the College Board and ETS through Contract PO # 0000004411, as well as Grant Award # 31-1992-701 from the United States Department of Education, Institute for Educational Sciences, as administered by the Temple University Laboratory for Student Success. Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the position or policies of the College Board, ETS, or the Institute for Educational Sciences, and no official endorsement should be inferred. The authors would like to gratefully acknowledge the AP Psychology and AP Statistics teachers who participated in this project. In addition, we would like to acknowledge William Disch for his efforts in assisting with item development; Jennifer Hedlund and Nan Taylor for assisting with teacher training; Cynthia Matthew for her assistance with data analysis; and James Freeman and the AP Psychology Readers for 2001 for allowing us to participate in their reading week.

Appendix A

Breakdown of open-response items on the augmented AP Psychology exam and their relative scale scores

Item number	Total points	Weighting value	Form A	Form B
51	1	2.50	2.5	
52	5	0.50	2.5	
53	4	0.63	2.5	
54	4	0.63	2.5	
55	4	0.83	3.3	
56	6	0.55	3.3	
57	2	1.65	3.3	
58	1	2.50	2.5	
59	2	1.25	2.5	
60	3	0.83	2.5	
61	4	0.63	2.5	
62	1	2.50	2.5	
63	2	1.25	2.5	
64	4	0.63	2.5	
65	3	0.83	2.5	
66	4	0.63	2.5	2.5
67	5	0.50	2.5	2.5
68	5	0.50	2.5	2.5
69	3	0.83	2.5	2.5
70	4	0.63		2.5
71	2	1.25		2.5
72	6	0.42		2.5
73	4	0.63		2.5

(*continued on next page*)

Appendix A (*continued*)

Item number	Total points	Weighting value	Form A	Form B
74	4	0.63		2.5
75	3	0.83		2.5
76	4	0.63		2.5
77	2	1.25		2.5
78	1	2.50		2.5
79	3	0.83		2.5
80	5	0.50		2.5
81	4	0.63		2.5
82	3	0.83		2.5
83	4	0.63		2.5
84	4	0.63		2.5
85	3	0.83		2.5
			49.9	50

Appendix B

Breakdown of open-response items on the augmented AP Statistics exam and their relative scale scores

Item number	Total possible raw score points	Weight value	Form A—maximum score after rescaling	Form B—maximum score after rescaling	Form C—maximum score after rescaling
51	8	0.416	3.33	3.33	3.33
52	8	0.416	3.33	3.33	3.33
53	8	0.416	3.33	3.33	3.33
54	5	0.336	1.68		
55	2	0.840	1.68		
56	3	0.560	1.68		
57	3	0.560	1.68		
58	5	0.666	3.33	3.33	
59	2	1.665	3.33	3.33	
60	1	3.330	3.33	3.33	
61	2	1.665	3.33	3.33	
62	2	1.665	3.33	3.33	
63	4	0.833	3.33		3.33
64	2	1.665	3.33		3.33
65	1	3.330	3.33		3.33
66	3	1.110	3.33		3.33
67	4	0.833	3.33		3.33
68	2	1.390		2.78	
69	1	2.780		2.78	
70	3	0.927		2.78	
71	2	1.390		2.78	
72	2	1.390		2.78	
73	3	0.927		2.78	
74	4	0.833		3.33	3.33
75	4	0.833		3.33	3.33
76	3	1.110			3.33
77	4	0.833			3.33
78	2	1.665			3.33
79	2	1.665			3.33
80	3	1.110		3.33	
			50.01	49.98	49.95

References

Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology, 87*, 559–575.

Blalock, H. M. (1972). *Social Statistics* (second ed.). New York: McGraw-Hill.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook i: Cognitive domain*. New York: Longmans Green.

Bond, T., & Fox, C. (2001). *Applying the rasch model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Cattell, R. B. (1971). *Abilities: Their strucure, growth, and action*. Boston: Houghton-Mifflin.

Ceci, S. J. (1996). *On intelligence (expanded ed.)*. Cambridge, MA: Harvard University Press.

Chubb, J. E., & Loveless, T. (Eds.). (2002). *Bridging the achievement gap*. Washington, DC: Brookings Institute.

College Board. (2004). *Exam Scoring*. Available from: http://apcentral.collegeboard.com/article/0,3045,152-167-0-1994,00.html.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fisher, W. P. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

Fordham, S., & Ogbu, J. U. (1986). Black students' school success: Coping with the "burden of 'acting white"'. *The Urban Review, 18*(3), 176–206.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York: Free Press.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jencks, C., & Phillips, M. (Eds.). (1998). *The black–white test score gap*. Washington, DC: Brookings Institution Press.

Linacre, J. M. (1988). *Facets: A computer program for many-facet rasch measurement (Version 3.3.0)*. Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet rasch measurement*. Chicago: MESA Press.

Linacre, J. M., Wright, B. D., & Lunz, M. E. (1990). *A facets model for judgmental scoring (MESA Memo No. 61)*. Chicago: MESA.

Luria, A. R. (1973). *The working brain*. London: Penguin.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests (Expanded ed.)*. Chicago: University of Chicago Press.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281–311.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in rasch measurement. *Structural Equation Modeling, 3*(1), 25–40.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629.

Stemler, S. E., Grigorenko, E. L., Jarvin, L., Macomber, D., & Sternberg, R. J. (2003a). Examining the utility of the theory of successful intelligence for enhancing the construct validity of the advanced placement psychology exam: Research Report submitted to the College Board.

Stemler, S. E., Grigorenko, E. L., Jarvin, L., Macomber, D., & Sternberg, R. J. (2003b). Examining the utility of the theory of successful intelligence for enhancing the construct validity of the advanced placement statistics exam: Research Report submitted to the College Board.

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences, 7*, 269–287.

Sternberg, R. J. (1985). *Beyond iq: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (1997). *Successful intelligence: How practical and creative intelligence determine success in life*. New York: Plume.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology, 3*, 292–316.

Sternberg, R. J. (2003). Culture and intelligence. *American Psychologist, 59*(5), 325–338.

Sternberg, R. J., & The Rainbow Project Collaborators (2005a). Augmenting the sat through assessments of analytical, practical, and creative skills. In: W. Camara & E. Kimmel (Eds.), *Choosing students. Higher education admission tools for the 21st century* (pp. 159–176). Lawrence Erlbaum Associates, Mahwah, NJ.

Sternberg, R.J., & The Rainbow Project Collaborators (2005b). The rainbow project: Enhancing the sat through assessments of analytical, practical, and creative skills. Manuscript submitted for publication.

Sternberg, R. J., The Rainbow Project Collaborators, & University of Michigan Business School Project Collaborators (2004). Theory based university admissions testing for a new millennium. *Educational Psychologist, 39*(3), 185–198.

Sternberg, R. J., Torff, B., & Grigorenko, E. L. (1998a). Teaching for successful intelligence raises school achievement. *Phi Delta Kappan, 79*, 667–669.

Sternberg, R. J., Torff, B., & Grigorenko, E. L. (1998b). Teaching triarchically improves school achievement. *Journal of Educational Psychology, 90*, 374–384.

Williams, B. (Ed.). (2004). *Closing the achievement gap: A vision for changing beliefs and practices*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.