Contemporary Educational Psychology 34 (2009) 195–209

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Using the theory of successful intelligence as a framework for developing assessments in AP physics

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

[a] Wesleyan University, 207 High Street, Middletown, CT 06549, United States
[b] Tufts University, Ballou Hall, Medford, MA 02155, United States
[c] Yale University, Child Study Center, 230 South Frontage Road, New Haven, CT 06519, United States

ARTICLE INFO

Article history:
Available online 5 May 2009

Keywords:
Intelligence
Achievement
Advanced placement
Physics
Cognitive processing

ABSTRACT

A new test of Advanced Placement Physics, explicitly designed to balance both content and cognitive-processing skills, was developed using Sternberg's theory of successful intelligence. The test was administered to 281 AP Physics students from 10 schools during the 2006–2007 school year. Six empirically distinguishable profiles of strengths and weaknesses emerged from an exploratory Q-type factor analysis across the four cognitive-skill areas assessed (i.e., memory, analytical, creative, and practical skills). These profiles replicated those found in previous research in the domains of AP Psychology and AP Statistics. Furthermore, achievement differences between ethnic groups on various cognitive subscales were reduced as compared with traditional estimates. The results provide evidence of the importance of integrating modern theories of cognitive processing into large-scale assessments.

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 futures (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). Indeed, many tests that serve as gatekeepers tend to emphasize only a limited range of skills (e.g., analytical and memory skills). Yet analytical and memory skills alone are not sufficient to succeed in the professional world. For example, although analytical skills are important to the physicist, who must compare and contrast competing explanations for phenomena and critically analyze data, other skills are important as well. It takes creative skills for the physicist to synthesize disparate findings and generate new theories, and practical skills to understand how theoretical findings may be used in the real world (e.g., to improve communication technology) as well as to persuade others of the value of the findings. To the extent that selection tests are weighted more heavily in favor of one particular type of skill, an entire professional field may suffer because it potentially will be dominated by individuals with a single profile of strengths and weaknesses, thereby inhibiting the capacity of the field to develop to its full potential. A balance of cognitive skills is important, regardless of one's professional domain. Thus, measurements should assess a broad profile of skills in students.

The aim of the current research was to examine the impact on student achievement of creating a set of modified, theory-driven examinations that expanded the range of cognitive skills assessed. The College Board's Advanced Placement (AP) program in Physics was used as a testing ground for the project.

1.1. The Advanced Placement Program

The College Board's Advanced Placement (AP) Program, initiated in 1955, was originally designed as a mechanism for granting exceptional high school students an opportunity for advanced study that would be equivalent to college-level programming. When this program began, it served only top students from a limited number of high schools, but in 2006, 666,067 graduating seniors (24% of all graduating seniors) at 16,000 secondary schools reported having taken at least one exam in one of the 37 courses across 22 subject areas offered by the AP program (College Board, 2007).[1]

[1] The courses offered by the AP Program are: Art History, Biology, Calculus AB, Calculus BC, Chemistry, Chinese Language and Culture, Computer Science A, Computer Science AB, Macroeconomics, Microeconomics, English Language, English Literature, Environmental Science, European History, French Language, French Literature, German Language, Comparative Government & Politics, US Government & Politics, Human Geography, Italian Language and Culture, Japanese Language and Culture, Latin Literature, Latin: Vergil, Music Theory, Physics B, Physics C, Psychology, Spanish Language, Spanish Literature, Statistics, Studio Art: 2-D Design, Studio Art: 3-D Design, Studio Art: Drawing, US History, and World History.

* Corresponding author.
E-mail address: steven.stemler@wesleyan.edu (S.E. Stemler).

0361-476X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cedpsych.2009.04.001

Each spring, students enrolled in AP courses are given the opportunity to take a high-stakes examination to demonstrate their mastery of the subject area. The exams are graded on a scale from 1 to 5, with a score of 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). Most colleges will grant credit to students scoring three or higher on the exam. Thus, the results of the test have important financial implications, as placing out of an introductory college courses could potentially save a student thousands of dollars in tuition in subsequent years. In addition, AP scores are frequently used in admissions decisions as predictors of college success (Morgan & Ramist, 1998). The limited number of chances to take the test, the potentially significant financial savings associated with the outcome, and the impact scores may have on college admissions decisions qualifies the AP examination as a high-stakes test that has a broad impact on hundreds of thousands of high school students each year.

Historically, the chief concern of AP exam developers has been with ensuring adequate content-area coverage. For example, the items on the AP Physics B exam are explicitly balanced to ensure proportionate representation of various subtopics within the domain of Physics (i.e., Newtonian mechanics; fluid mechanics and thermal physics; electricity and magnetism; waves and optics; and nuclear physics). Traditionally, however, there has been no systematic attempt explicitly to balance items for the cognitive-skill areas they assess.

1.1.1. Ethnic differences in achievement

One of the biggest challenges facing the AP program is in the recruitment of minority students to participate in the program. Research has found that African–American and Latino students enroll in AP courses at approximately half the rate of White students. In particular, minority students enroll in AP math, science, and English classes at lower rates than White students at comparable schools (Klopfenstein, 2004; Ramist, Lewis, & McCamley-Jenkins, 1994). As a result of this differential enrollment, fewer minority students end up taking AP exams. In 2006, approximately 21% of all students who took one or more exams were African–American or Latino; by way of comparison, approximately 30% of students enrolled in high schools were African–American or Latino (College Board, 2007). Because taking an AP course is a strong predictor of whether a student will take an upper level class or major in that subject in college (Dodd, Fitzpatrick, DeAyala, & Jennings, 2002; Morgan & Maneckshana, 2000), the AP courses that students choose to take have important implications for their future course of study and, eventually, their profession.

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 across ethnic groups (Chubb & Loveless, 2002; Jencks & Phillips, 1998). Indeed, research suggests that White students receive higher scores on standardized tests than African–American, Latino, and Native-American students as early as preschool (Nettles & Nettles, 1999). This difference is dramatic on most conventional achievement tests; nearly a full standard deviation separates the average scores of African–American and White high school students (Hedges & Nowell, 1998). This pattern holds for scores on the AP exam as well. For example, in 2006 the mean score for African–American test-takers across all AP exams was 1.96, compared with 2.96 for White students (College Board, 2007). This difference is not only large but consequential: because three is typically a passing score for getting college credit, the average White student will "pass" a given AP exam while the average African–American student will "fail" it.

The difference in scores of students from different ethnic backgrounds is more dramatic in some domains than in others. For example, there is little difference between the scores of White students and African–American students on the AP Studio Art: 3D-Design exam; the average score of African–American students was 2.68 compared with 2.95 for White students. But a difference of 1.13 separates the average scores of White students and African–American students on the AP Physics C exam; comparable results are 1.15 for AP Microeconomics, and 1.35 for AP Computer Science scores (College Board, 2007). As AP scores are a useful indicator of college success and an important consideration in the college-admissions process, differences in these scores have high-stakes consequences.

Researchers have proposed several possible reasons for the achievement gap between White students and underrepresented minorities, including genetic differences (Herrnstein & Murray, 1994), cultural differences (Fordham & Ogbu, 1986; Williams, 2004), social–psychological differences (Steele, 1997), and differences in the quality of instruction (Hanushek & Rivkin, 2006). Another potential reason for this persistent difference, however, is that traditional achievement tests have assessed a fairly limited range of cognitive processes, ignoring other important skills.

Sternberg and colleagues have demonstrated in a series of studies that when assessments are designed in such a way that they expand the range of cognitive skills assessed, the achievement gap between White students and minority students is reduced. For example, in a recent study designed to create assessments that would enhance the predictive power of the SAT, Sternberg and The Rainbow Project Collaborators, (2006) found that adding assessments of creative and practical skills doubled the power of the battery to predict first-year college GPA compared with the use of the SAT alone. In addition, differences in achievement between White and African–American students were reduced on measures of creative skills, and differences in achievement between White and Latino students were reduced on assessments that emphasized practical skills and creative skills.

The decrease in the achievement gap as a result of measuring a broader range of cognitive skills has also been demonstrated in the context of the AP program. Stemler, Grigorenko, Jarvin, and Sternberg (2006) designed augmented versions of the AP Psychology and AP Statistics examinations that included practical and creative subscales. A key finding was that the effect-size difference between African–American students and White students was virtually nonexistent for both the creative subscale ($d = -0.02$) and the memory subscale ($d = 0.04$) of the modified exams. The largest difference between Latino students and White students was observed on the memory subscale of the modified AP Psychology exam, in which 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 ($d = -0.13$). Results of the AP Statistics exam showed a similar pattern. Overall, the findings from these past studies suggest that developing assessments that measure a broad range of cognitive abilities may help to create more equitable achievement tests.

1.2. Theoretical framework

In recent years, designers of large-scale testing programs, recognizing the important social, economic, and ethical consequences associated with standardized testing, have become particularly interested in linking educational assessment to modern theories of cognitive-processing skills (Embretson & Reise, 2000; Irvine & Kyllonen, 2002). Capitalizing on this idea, the current project involved the development of an augmented test in the subject area of AP Physics B that was explicitly linked to Sternberg's (1997,

1999) theory of cognitive-processing skills. Many tests that have made an attempt explicitly to measure students' cognitive skills have either explicitly or implicitly used Bloom's (1956) taxonomy of the cognitive domain as their theoretical basis (e.g., Garden et al., 2006; Martin et al., 2008; Mullis et al., 2008). Within this framework, students' intellectual skills are thought to proceed along a linear hierarchical path, beginning with knowledge of a topic, progressing through the stages of comprehension, application, analysis, synthesis, and evaluation. Knowledge is demonstrated through the recall of information. Comprehension is demonstrated when an individual can restate a problem in his/her own words. Application is demonstrated when an individual can apply what s/he has learned to a novel situation. Analysis is demonstrated when an individual can separate a problem into component parts or see an underlying structure to a problem. Synthesis occurs when an individual puts separate parts together to form a new whole. Evaluation is demonstrated when an individual makes judgments about the value of ideas (Clark, 1999). Thus, if an individual is able successfully to analyze a problem, it is assumed by users of the taxonomy that the individual will be able to apply his or her knowledge of the topic as well.

A more recent theory of cognitive processing is Sternberg's theory of successful intelligence. According to the theory (Sternberg, 1984, 1985, 1997, 1999), a common set of processes underlies all aspects of problem solving. These processes are hypothesized to be universal. 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 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 processes 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. *Memory* skills are also foundational for each type of thinking.

Construct validation of Sternberg's theory has been described elsewhere and only can be summarized here. Some of the main findings from these studies are the following:

1. The analytical, creative, and practical aspects of intelligence can be measured via both multiple-choice and essay formats. Structural-equation modeling provides some support for this model of intelligence over competing models, such as a model of an overarching general factor and a model of content factors (Sternberg, Castejón, Prieto, Hautamäki, & Grigorenko, 2001; Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999).
2. Tests of analytical intellectual abilities as measured componentially (with decomposition of reaction times) tend to correlate well with conventional tests of intellectual abilities, because these tests measure what the conventional tests measure (Guyote & Sternberg, 1981; Sternberg, 1980, 1983; Sternberg & Gardner, 1983).
3. Tests of creative intellectual abilities are relatively domain specific, correlating weakly to moderately with conventional tests of intelligence, with the correlations higher the more nonentrenched the content of the conventional tests (Sternberg, 1982; Sternberg & Gastel, 1989a, 1989b; Sternberg & Lubart, 1995).
4. Tests of practical intellectual abilities correlate weakly with conventional tests of intelligence and predict real-world occupational success as well as or better than conventional tests of academic intelligence (Sternberg, Wagner, & Okagaki, 1993; Sternberg, Wagner, Williams, & Horvath, 1995; Sternberg et al., 2000; Wagner, 1987; Wagner & Sternberg, 1985), thus complementing conventional tests. Under special circumstances, tests of practical intelligence may show negative correlations with conventional ability tests (Sternberg et al., 2000).

The two theories of cognitive processing (Bloom's and Sternberg's) are not entirely incompatible. Indeed, the *application* level of Bloom's taxonomy is quite similar to the *practical skills* articulated by Sternberg's theory. In addition, the *synthesis* level of Bloom's taxonomy shares some common features with the *creative skills* aspect of Sternberg's theory. The key distinction between Sternberg's theory and Bloom's is that whereas Bloom's taxonomy specifies a hierarchical progression of cognitive skills, Sternberg's theory takes an interactive and profile-oriented approach. Thus, the theory of successful intelligence suggests that is possible for one person to have high levels of practical skills and low levels of creative and analytical skills, whereas another person may have high levels of creative skills and low levels of practical and analytical skills.

A key advantage to using an expanded theory of cognitive-processing skills in test construction is that it can provide more useful information about and for individual students. Within the context of such a paradigm, 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, students 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. Prior research also has shown that traditionally underrepresented minority students stand especially to benefit from broader measures of cognitive skills (Sternberg & The Rainbow Project Collaborators, 2006; Sternberg, Torff, & Grigorenko, 1998a, 1998b).

The theory of successful intelligence suggests that some students may be more capable of demonstrating their knowledge when problems are placed within a practical but not an analytical context. Others may show the reverse pattern. For example, Nuñes and colleagues (Carraher, Carraher, & Schliemann, 1985; Nuñes, 1994; Nuñes, Schliemann, & Carraher, 1993) studied Brazilian children who, for economic reasons, often worked as street vendors and had very little formal schooling. These children were successful in completing trade-related computational operations but were unable to solve computationally similar problems when they were presented in abstract terms. Conversely, many schoolchildren could solve paper-and-pencil arithmetic questions but could not solve the same type of problem in a different applied context (Perret-Clermont, 1980). Similar results have been found in the literature on logical reasoning (Leighton & Sternberg, 2004; Sternberg & Ben-Zeev, 2001).

1.3. Present research

The purpose of the current research was to examine the impact on performance of creating assessments in the domain of Physics

that were based on Sternberg's construct-validated theory of cognitive-processing skills described above (Sternberg, 1997, 1999). We were particularly interested in examining the following research questions:

1. Is it possible to develop items that tap a broad range of cognitive skills (memory, analytical, creative, and practical) and for those items then to be reliably coded by subject-matter experts and experts on cognitive processes into the appropriate category?
2. Is it possible to create an integrative test of cognitive-processing skills and domain knowledge in the area of physics that demonstrates desirable psychometric properties?
3. Does assessing a broad range of cognitive skills within the context of quantifying domain knowledge reduce ethnic differences in achievement when compared with conventional assessments?
4. Do students show uneven profiles of strengths and weaknesses across different cognitive skills or is there a relatively uniform pattern of performance across skills? If there are profile differences, are those differences systematically related to sex or ethnicity?

2. Study 1

The first study involved the development of the items for the augmented examination and the systematic evaluation of the content validity of the test. This study involved three components. In the first component, 10 high school and college physics teachers (subject-matter experts) were educated in how to develop items that were balanced for content and process knowledge. In the second component, three independent subject-matter experts rated the full set of developed items based on how much of each type of thinking process was required to answer the question. Finally, in the third component, two process experts rated the final selection of items based on how much of each type of thinking they required. We explicitly relied on participation of high school and college physics teachers to (i) ensure the proper content coverage of AP Physics issues and (ii) replicate, as much as possible, the style of item development utilized for real AP examinations.

2.1. Method

In consultation with the College Board, invitations were sent to practicing AP teachers and readers to attend a 3-day summer workshop at Yale University. Ten participants were selected: six men and four women. Half of them were high school teachers of AP Physics, and the other half taught at the college level. They came from seven different states, and represented a range of experiences, having taught for 9–38 years, and having participated as AP exam readers for 2–7 years. The ten selected teachers were provided with lodging and meals, educational materials, a tour of the Yale campus, and financial compensation for their time.

The goal of the workshop was to teach the participating subject-matter experts to design items that, in addition to content, explicitly

Fluid Mechanics and Thermal Physics
Memory Item
On a pressure-volume graph, an isobaric process is represented by what shape graph?
a. a diagonal lime
b. a horizontal line
c. a vertical line
d. a parabola
e. a hyperbola
Letter of Correct Answer: b

Waves and Optics
Analytical item
Two waves traveling in opposite directions pass through each other in the ocean. The first wave has an amplitude of 1 meter while the second wave has an amplitude of 2 meters. What is the resultant amplitude of the highest crest of the superimposing waves?

Electricity and Magnetism
Creative item
You have at your disposal a ruler, a lamp (with a 60 W bulb), a meterstick, and ammeter, a volt meter, and a solar cell of unknown efficiency. It is a sunny day. Devise a method to measure the total power output by the sun. Would you expect your answer to be an overestimate or an underestimate? Why?

Newtonian Mechanics
Practical item
Your car is stuck in a snowbank. What can you do to get it out?
I. Throw sand under the tires to increase the friction between the tire and the ground.
II. Remove items from the car to reduce the friction between the tires and the ground
III. Add weight over the wheels to increase the friction between the tires and the ground.
a . I only
b. II only
c. I and II
d. I and III
e. III only
Letter of Correct Answer: d

Fig. 1. Example items.

measured memory, analytical, creative, and practical skills. Item development roughly followed the design framework proposed by Perret-Clermont (2004), in which item development is "guided by four critical questions: (a) What does it mean to know and to inquire? (b) What constitutes evidence of knowing? (c) How can that evidence be elicited from students? (d) What are appropriate techniques for making valid inferences about what students know, from what students do?" (pp. 3–4). The item-development team members were instructed to follow the same approach to test construction used by the actual AP program; however, they were given more leeway in developing additional open-response items. During the practice session, which unfolded over 2 days, participants were asked to develop assessment items and corresponding rubrics and then reviewed other participants' items. The subject-matter experts were then asked to develop as many new items as possible over the course of the three months subsequent to the workshop and to submit the items to the project team for review.

During the second step of the study, three independent subject-matter experts were asked to rate a total of 264 newly developed AP Physics items based on the degree to which each item tapped the following skills: (i) memory, (ii) analytical, (iii) creative, and (iv) practical skills. Fig. 1 presents examples of each of the different item types.

Each judge was asked to rate the percentage of thinking in each item that required each skill, such that the total percentage would always add to 100%. For example, a given item may have tapped 20% memory skills, 40% analytical skills, 10% creative skills, and 30% practical skills. The judges were also asked to assign each item to one of the five content areas covered on the actual AP Physics test (i) Newtonian mechanics, (ii) fluid mechanics and thermal physics, (iii) electricity and magnetism, (iv) waves and optics, and (v) atomic and nuclear physics, and rate (1 = poor to 5 = great) each of the newly developed items in terms of the following dimensions: (i) content area fit; (ii) quality of the item, and (iii) difficulty of the item. In addition, the judges were asked to evaluate the keyed responses and the distracters for each item.

Next, the items were reviewed by our own internal psychometric team for common errors in item development (see Gronlund, 2002) and for redundancy with regard to content coverage. Because the ultimate goal was to develop a pilot test that would look and feel as close as possible, logistically, to the actual AP exam (particularly in terms of amount of time allocated to take the test), the project team selected the most technically sound set of items from the initial item pool and subjected them to pilot testing. The pilot test was designed to be relatively balanced by both content area and cognitive-skill areas assessed and contained items that passed through external review for common errors in item development. In the end, a total of 73 items were used in the pilot test (53 multiple-choice and 20 open response). These items were then administered to a sample and the psychometric data analyzed. Based on the results of the pilot testing, several multiple-choice items were eliminated and some new items were developed to replace pilot items with poor psychometric properties. The final version of the test consisted of 69 items (45 multiple-choice and 24 open-response).

Finally, two process experts who were well-versed in the theory of successful intelligence then rated the extent to which each of 69 final items tapped (i) memory, (ii) analytical, (iii) creative, and (iv) practical skills. Again, the percentages of each type of thinking required by an item were expected to add up to 100%.

2.2. Results

A total of 264 new items were developed during Study 1 by the teachers on the test-development team. The test-development team was asked to develop items related to the five subdomains covered on the actual AP Physics test (i) Newtonian mechanics, (ii) fluid mechanics and thermal physics, (iii) electricity and magnetism, (iv) waves and optics, and (v) atomic and nuclear physics. The percentage of items receiving the top score of five in terms of content area fit (i.e., great content area fit), as rated by three independent AP Physics teachers, ranged from 72% to 84% across the three raters (median = 76%).

Table 1
Results of interrater-reliability analyses related to cognitive demand of items.

Cognitive process	Raters 1 and 2	Raters 1 and 3	Raters 2 and 3
Memory			
Percent agreement[a]	0.56	0.51	0.45
Cohen's kappa[a]	–	–	−0.10
Analytic			
Percent agreement	0.89	0.70	0.65
Cohen's kappa	0.03	0.20	0.09
Practical			
Percent agreement	0.65	0.60	0.63
Cohen's kappa	0.30	0.17	0.26
Creative			
Percent agreement	0.80	0.71	0.78
Cohen's kappa	0.45	0.24	0.36

Raters 2 and 3 shared $N = 223$ ratings.
Note: $N = 264$ total items.
[a] Raters 1 and 2 and Raters 1 and 3 shared $N = 205$ ratings.

The same three subject-matter experts were also asked to rate the extent to which each item required each of the four cognitive processes in order to be correctly answered.[2] Table 1 presents the results from the interrater-reliability analyses.

Because we were interested in determining the extent to which raters could agree on the exact categorization of items into content areas and cognitive processes, and each category was qualitatively different, the most useful estimate of interrater reliability is a consensus estimate (e.g., percent agreement or Cohen's kappa). The simplest measure of consensus is the percent agreement statistic; however, the problem with percent agreement is that it can be artificially inflated if there are low base rates. Consequently, an alternative statistic is Cohen's kappa, which corrects for agreement by raters that would be attributable to chance alone.[3]

[2] The raters gave numerical responses ranging from 0 to 100 for each process, thereby facilitating the computation of consistency estimates of interrater reliability. For example, Rater 1 might have felt that Item 1 demanded of the test-taker 60% memory skills, 5% analytic skills, 0% creative skills, and 35% practical skills. Thus, for each item, the percentages assigned to the four processes had to add to 100%. These percentages were then recoded to make them categorical in order to facilitate the computation of consensus estimates of interrater reliability. For the latter purpose, the data were recoded so that each item was coded as either "requiring" or "not requiring" each of the four cognitive processing skills (e.g., practical skills). Thus, if a rater assigned any value greater than 0 to a cognitive processing component, the item was coded as requiring that component. Percent agreement was then computed by evaluating the agreement between raters on the number of items requiring each cognitive processing component.

[3] Although the interpretation of the percent agreement statistic is intuitive and straightforward, some caution must be exercised in interpreting Cohen's kappa. Kappa was intended to describe the extent to which raters agree over and above the degree to which they would be expected to agree by chance alone. Some authors have suggested guidelines for interpreting kappa (e.g., Landis & Koch, 1977; Stemler & Tsai, 2008); however, other authors (Krippendorff, 2004; Uebersax, 2002) have argued that the kappa values for different items or from different studies cannot be meaningfully compared unless the base rates are identical. Consequently, these authors suggest that although the statistic gives some indication as to whether the agreement is better than that predicted by chance alone, it is difficult to apply rules of thumb for interpreting kappa across different circumstances. Instead, Uebersax (2002) suggests that researchers using the kappa coefficient look at it for up or down evaluation of whether ratings are different from chance, but that they not get too invested in its interpretation. Readers seeking a detailed guide for calculating and interpreting the most commonly reported interrater reliability statistics are encouraged to consult Stemler (2004) and/or Stemler and Tsai (2008).

Table 2
Number of items categorized as primarily (>51% of skill tapped) falling into each domain.

Rater	Memory	Analytic	Practical	Creative	No primary theme	Total *N* of Items with primary theme
Rater 1	27	94	0	57	86	178
Rater 2	11	130	20	49	54	210
Rater 3	43	87	57	66	11	253

Percent agreement statistics were computed between each pair of subject matter expert raters for each cognitive domain. Overall, the results of the percent agreement analyses look promising within the context of an exploratory research study. In particular, Raters 1 and 2 showed consistently high levels of agreement. The memory subscale had the lowest levels of agreement across all raters (median = 0.51) whereas the creative subscale exhibited the highest levels of agreement (median = 0.78). These findings were somewhat surprising as one might expect higher agreement on memory items; however, the high levels of agreement on the creative items may be due to the relative novelty of this kind of item, at least from the perspective of the subject-matter experts. In general, the pattern of agreement ratings ranged from the .50s to the high .70s. Table 2 presents the number of items each rater classified as primarily belonging to one of the four cognitive processing areas.

A pilot version of the exam was constructed and the pilot test was administered between March and May of 2006. A total of seven teachers from seven different schools participated in the pilot study and a total of 138 students took the pilot augmented version of the AP Physics exam. Of the students who took the exam, there were 79 males, 56 females, and three students who failed to indicate their sex. In terms of ethnicity, 71 students were White, 56 were Asian–American, four students were African–American, and one student reported multiple ethnicities, with the remaining six students not responding.

Based on the analyses of these pilot data, items were revised (e.g., open-ended question prompts that had initially yielded few correct answers were clarified) and test length was optimized (e.g., noting that the frequency of responses declined sharply after multiple-choice item 45, we decided to reduce the number of multiple-choice items from 53 to 45). A total of eight multiple-choice items that were included on the pilot test were cut from the main test. Of those eight items, four represented Newtonian mechanics, two represented fluid mechanics and thermal physics, and two represented electricity and magnetism. In terms of cognitive processes tapped by the items, two were tapping memory skills, five were tapping analytic skills, and one was tapping practical skills. In addition, a subject matter expert was retained to develop substitute items in order to ensure a balanced representation of memory, analytical, creative, and practical items. We thus mimicked the initial item development phase by relying on a subject matter expert to develop items that were then reviewed and revised in collaboration with the process experts.

In selecting the total number of items to be used on the final version of the exam, we had three major criteria: (i) the final item set needed to represent a balance of both content and cognitive-processing skills; (ii) the final item set was to consist of items that demonstrated desirable psychometric properties (e.g., item difficulty and discrimination values) from the pilot test, and (iii) the final item set was designed to take approximately 1.5 h to complete, in order to remain comparable in length to the actual AP Physics exam. Consequently, we were less concerned with ensuring an even number of items on the test and more concerned with choosing the best possible item set that met the aforementioned criteria. Similarly, during the pilot testing phase, our primary aim was to pilot test as many items as possible under the given time constraints so as to have a large pool of items to select from in developing the final version of the test.

The items on the final version of the exam were then rated by two independent cognitive-processing experts with regard to their cognitive-skill classification. Mimicking what was done for the subject matter expert raters, the two processing expert raters specified the percentage of each cognitive skill area demanded by each item. In order to facilitate the computation of consensus estimates these ratings were recoded so that each item had one primary cognitive processing skill (i.e., a rater would have assigned it a value greater than 50%). The process raters exhibited very high consensus estimates of reliability on the memory subscale (agreement = 98%, kappa = .85), strong exact agreement on the analytic subscale (agreement = 82%, kappa = .60), and high levels of agreement on the creative subscale (agreement = 91%, kappa = .58) and practical subscale (Agreement = 85%, kappa = .62).

The ratings of experts in cognitive processing of the final 69 items were found to be highly reliable and yielded more variability than those of subject-matter experts. As such, their classifications were used to construct the process subscales that are employed in future analyses. Although the ratings of these process experts exhibited moderately high agreement, to be conservative, only those items on which both raters agreed on the categorization of the primary cognitive process were included in each subscale. For example, only if both raters agreed that the dominant process involved in a question was creative would that item then be included in the creative subscale. Of 69 items, 53 met this criterion––36 were multiple-choice items and 17 were open-response items. Table 3 presents a Table of Specifications for the complete test of 69 items as well as the subset of 53 items meeting the aforementioned criteria broken down by content area, item type (MC v. OR), and cognitive process.

After the items had been revised and re-rated, we then proceeded with the execution of Study 2.

3. Study 2

Study 2 consisted of three parts: (i) an examination of the internal psychometric properties of the full augmented exam, (ii) an analysis of ethnic differences in achievement by subscale, and (iii) a Q-factor analysis of student responses to explore the existence of different achievement profiles based on cognitive-skill areas.

3.1. Method

3.1.1. Sample

In order to select the teachers, we acquired from the College Board a list of all 2383 teachers scheduled to administer the AP Physics B exam in the spring of 2007, divided into nine geographical regions. Next, we took the existing AP Physics teacher database and added demographic information for each school. The demographic information included the absolute numbers of African–American, Latino, and White students enrolled in the school, based on school information obtained from the National Center for Education Statistics (http://nces.ed.gov/ccd/schoolsearch/). For each region, the schools were ranked in decreasing order of number of African–American students and Latino students. A list was created whereby half of the teachers to be recruited in the region were obtained from the schools with the most African–American students,

Table 3
Table of specifications for 69 items developed for the new AP physics test.

Content domain	Cognitive process									
	Memory		Analytic		Creative		Practical			
	MC	OR	MC	OR	MC	OR	MC	OR	Total *N* of items	Alpha
1. Newtonian mechanics	0	0	5	3	4	1	4	1	18	0.55
2. Fluid mechanics and thermal physics	3	0	4	2	0	1	1	1	12	0.50
3. Electricity and magnetism	2	0	2	5	3	0	3	2	17	0.64
4. Waves and optics	0	1	3	1	2	0	3	2	12	0.07
5. Atomic and nuclear physics	0	0	3	2	2	0	1	2	10	0.58
Total *N* of items	5	1	17	13	11	2	12	8	69	
Cronbach's alpha for scale	0.42		0.65		0.39		0.54		0.78	
Table of specifications for 53 items used in subsequent analyses										
1. Newtonian Mechanics	0	0	2	2	2	0	4	1	11	.38
2. Fluid Mechanics and Thermal Physics	3	0	3	1	0	1	1	1	10	.51
3. Electricity and Magnetism	1	0	2	4	2	0	2	2	13	.61
4. Waves and Optics	0	0	3	1	2	0	3	1	10	.06
5. Atomic and Nuclear Physics	0	0	3	2	2	0	1	1	9	.57
Total *N* of items	4	0	13	10	8	1	11	6	53	
Cronbach's Alpha for Scale	.52		.59		.30		.43		.76	

MC = Multiple-choice item; OR = Open-response item.

while the other half of the teachers to be recruited were from the schools with the most Latino students.

In the first wave of recruitment, we sent invitation letters to 35% of teachers in each region, for a total of 62 teachers. Within each region, we randomly selected half the potential participants among teachers who had taught for 0–5 years, and half among teachers who had taught for 6 years or more. For example, in the *Northeast* region, there were a total of 358 teachers. We randomly selected 10 teachers to send invitations to, five of whom had taught for 5 years or less, and five of whom had taught for 6 years or more.

Ultimately, a total of 10 teachers from 10 different schools participated in the study and a total of 281 students took the augmented version of the AP Physics exam. Of the students who took the exam, there were 151 males, 104 females, and 26 students who failed to indicate their sex. In terms of ethnicity, 87 students were White, 69 were Asian–American or Pacific Islanders, 40 were Latino, 32 were African–American, and five reported multiple ethnicities, with the remaining 48 students not responding.

3.1.2. Procedure

Because teachers administered the test to their students at their own pace, it is difficult to say with certainty what procedures were followed. All teachers were asked to administer the exam between March and May of 2007 as a practice exam for their students in preparation for the actual AP exam in May. Participating teachers were offered $150 for administering the augmented exam and were provided with the answer key immediately after exams were returned via FedEx.

As with the actual AP exam, students were expected to complete the multiple-choice section of the exam first, followed by the open-response section. The total test was estimated to take 1.5 h. After completing the assessment, students were asked to fill out a questionnaire that asked them to indicate their ethnicity, sex, whether they owned a cell phone (an indicator of socioeconomic status), current grade in physics, SAT scores, how much they liked physics, how many other AP classes they had taken, and the number of hours they studied physics each night. A total of 13 students did not complete the multiple-choice section, 15 students did not complete the open-response section, and 20 students did not complete the questionnaire. It is not clear whether this is a reflection of the individual students or their teachers. These students were not excluded from the analyses, but information was coded as missing where appropriate. Research assistants at Tufts University scored the multiple-choice section of the exams. Open-response sections were scored by an independent rater who was provided with scoring guidelines. A second rater scored half of the exam; the raters showed high agreement (Agreement = 90%).

3.2. Results

First, classical item statistics (e.g., item-difficulty estimates and item-discrimination values) are presented and discussed. These statistics allow one to determine how well each item on the exam functioned––whether it was of the appropriate difficulty level and to what extent it was able reliably to discriminate between people of different ability levels. Next, the internal-consistency reliability and the results of a Rasch analysis (e.g., item maps, item-difficulty estimates, and fit statistics) are reported; these results provide important evidence as to how the exam functioned as a whole. The item- and test-level results address the research question of whether it is possible to create a psychometrically-sound test grounded in a theory of cognitive processing in physics. Finally, information about how different groups performed on the test is presented. This information includes analyses of differential test functioning based on ethnicity (to address the impact of testing a range of cognitive processes on the achievement of different demographic groups), and Q-type factor analysis (to determine if individuals exhibit varied profiles of strengths and weaknesses across cognitive skills).

3.2.1. Item statistics

Of 45 multiple-choice items, 11 had item difficulty values less than .30, indicating that fewer than 30% of the participants answered those items correctly. Six items had difficulty values greater than .70, indicating that more than 70% of participants answered those items correctly. Taken together, this means that just over half of the multiple-choice items on the test (58%) fell within the standard target item-difficulty range of .30 to .70, indicating that the items on the multiple-choice section of the augmented exam exhibited a wide range of difficulty levels. There were five items on the open-response section of the AP Physics exam, all with multiple sections and many with multiple subsections, leading to a total of 24 separately analyzed items. Each of the 24 items was scored on a partial-credit scale that ranged from 2 to 6 points per item. The average score per item ranged from 0.19 to 2.89.

Item-total correlations were computed for all items by correlating scores on each individual item with the raw sum of scores on all of the multiple-choice and open-response items on the test.

Two items had slight negative item-discrimination values (Item 26 $p = .15$, $r = -.01$; Item 40 $p = .18$, $r = -.05$), indicating that performance on these items was negatively related to overall performance on the test. The scoring key was checked and the answer options to these items reviewed and all were found to be accurate. Consequently, because the item-discrimination values were close to zero, they were not a major concern or justification for eliminating the items. The remaining items all had positive item-discrimination values, indicating that performance on each item was positively correlated with performance on the remainder of the exam.

3.2.2. Test statistics

3.2.2.1. Reliability. The internal-consistency reliability for the multiple-choice section (Cronbach's alpha = .64), open-response section (Cronbach's alpha = .72), and total test (Cronbach's alpha = .75) were all acceptable within the context of an exploratory study. Although it would be preferable to see higher reliabilities for a high-stakes test, it is not surprising that the internal-consistency reliability may have been reduced by the addition of a second underlying dimension to the test (i.e., cognitive-processing skills), which reduces test homogeneity.

3.2.2.2. Rasch measurement statistics. The Rasch model offers a way to understand item responses as a function of person and item parameters (Rasch, 1960/1980; Wright & Stone, 1979). Under this model, the probability of any given correct response is modeled as a logistic function of the difference between person ability and item difficulty. Thus, the probability of a correct response will be higher as the person's ability increases relative to item difficulty. The following statistics all address the degree to which this assessment conforms to the expectations of the Rasch model. Because the data used in this study combined both dichotomous response data found in the multiple-choice section of the exam as well as partial credit data found in the open-response section, a partial credit Rasch model was used (Wright & Masters, 1982).

The Rasch model is best used when there is one dominant underlying dimension of interest being measured. In this study, that overarching dimension is general Physics knowledge. Indeed, we are not arguing that someone with high creative processing skills but no physics knowledge would do well on the creative physics items. Rather, content mastery is a necessary pre-requisite to performance. Consequently, the novel aspect of the current data is that certain types of items may allow students to express their mastery of that content in different ways, and it is for this reason that a Rasch approach was used.

As can be seen in Fig. 2 below, the items on the exam were somewhat difficult relative to the abilities of the test-takers. The mean person ability estimate was −0.41 ($SD = 0.46$) logits, indicating that the sample of test-takers had ability levels that were slightly below the average difficulty levels of the items; an ideal solution would find a mean person ability estimate of 0.00. More items within the −1.00 to 0.00 logit range would be useful in discriminating among the group of test-takers who were clustered at

```
TABLE 12.12 AUGMENTED AP PHYSICS EXAM              ZOU923WS.TXT Mar 16 14:38 2008

INPUT: 281 persons  69 items  MEASURED: 281 persons  69 items  199 CATS     3.61.1
--------------------------------------------------------------------------------

       persons MAP OF items
               <more>||<frequ>
    3                ++
                     ||
                     ||
                     ||
                     ||
                     ||  C_38
    2                ++
                     ||  A_5
                     ||T 42       C_44
                     ||  20
                     ||  43       A_P2Q2
                     ||
    1                ++  P_24     P2Q1
                  .  ||S A_11     31
                  #  ||  22       P5Q1
                  # T||  C_17     P_21     M_25     M_27     A_28   A_P1Q1.2  P_P1Q3
                .##  ||  A_19     A_23     A_35     P_39     P_40   6         A_P4Q1
              .####  ||  M_14     P_33     36       P_P3Q3.2 P5Q4
    0       .####### S++M M_37   41       P_P2Q4   P3Q1.2   P_P3Q4
            .#######  ||  P_10     A_2      A_45     7        P4Q2
        ########### M||  P_1      32       C_P4Q3
       .###########  ||  C_13     P_16     A_18     A_P1Q1.1 P2Q3   P_P4Q4   P_P5Q3
         .#########  ||  P_15     P1Q2     P5Q2
            .###### S||S A_29     C_4      C_8      A_9
   -1           .##  ++  C_34     A_P3Q1.1 A_P3Q2.1 A_P3Q2.2
                 .#  ||  P1Q4
                 .# T||  A_12     C_30
                 .#  ||  A_26     P_3
                  .  ||T
                     ||
   -2              .  ++
                     ||
                     ||  A_P3Q3.1
                     ||
                     ||
                     ||
   -3                ++
               <less>||<rare>
 EACH '#' IS 4.
```

Note: Items preceded by an M_ correspond to items tapping Memory skills; Items preceded by an A_ correspond to items tapping Analytical skills; Items preceded by a C_ correspond to items tapping Creative skills; Items preceded by a P_ correspond to items tapping Practical skills.

Fig. 2. Item map of the augmented AP physics test.

this ability level. The item map further shows that cognitive processes were relatively independent of item difficulty. That is, it was not the case that all items that corresponded to a particular process (i.e., analytical items) were harder or easier than other items; rather, items on the same subscales were spread out in terms of difficulty level.

All items except for Problem 4, Question 4, and Problem 2, Question 4, had infit mean square values that fell within the desired range of 0.70–1.30 (Bond & Fox, 2001). Problem 4, Question 4, and Problem 2, Question 4 had infit mean square values of 1.45 and 1.70, respectively, indicating people who were expected to answer these items correctly answered incorrectly and vice versa. Not surprisingly, these were the same two items that exhibited negative discrimination values. The Rasch model is probabilistic, not deterministic, so it is consistent with some low-ability people answering difficult items correctly and some high-ability people answering easy items incorrectly; but the generally good infit statistics suggest that, in most cases, this did not happen more than was predicted by the model. Furthermore, the fit statistics demonstrate that the data were a good fit to the unidimensional Rasch model.

Next, the Rasch model offers person and item reliability estimates that can be interpreted in the same way as Cronbach's alpha statistic. However, the Rasch reliability estimates take into account the accuracy with which the underlying construct is measured, giving more weight to those items or people that provide better measures of the construct (Bond & Fox, 2001). By this model's specifications, the person reliability estimate was .75, with a separation value of 1.71. The reasonably high reliability estimate means that, given a similar test, the order of test-taker abilities would likely stay the same. The separation value was below the typically accepted value of 2.00, indicating that the students taking the exam were not well spread out with regard to their ability (Bond & Fox, 2001). That is, the overall ability of the test-takers was clustered together in such a way that few of the items on the exam were useful in discriminating among these test-takers. As indicated by the mean person ability estimate, these test-takers were clustered within the −1.00 to 0.00 logit range.

In contrast, the item separation was 6.64, with an item reliability estimate of .98. The high item-separation estimate indicates that the items were spread out across a range of difficulty levels and the high item reliability suggests that given a different sample of test-takers, the order of item difficulties would likely stay the same.

Overall, the Rasch measurement statistics revealed that the items on the test followed a linear progression in terms of difficulty that was spread out across all items on the test. The majority of items conformed well to the expectation of the model that test-takers are able to correctly answer questions at or below their ability level. However, the students were not well spread out in terms of their abilities so there were a limited number of items that served as ideal measures of ability for this particular sample of test takers.

3.2.3. Group statistics

The test's functioning for different groups was assessed looking at ethnic groups and cognitive profile groups.

3.2.3.1. Ethnic differences in ability. As reported previously, of the 281 students who took the exam, 87 were White, 69 were Asian–American or Pacific Islanders, 40 were Latino, 32 were African–American, five reported belonging to multiple ethnic groups, and the remaining 48 chose not to report their ethnicity. Individuals who did not report their ethnicity or reported affiliating with multiple ethnic groups were excluded from these analyses, so the data were split into four groups based on the ethnicity reported by the participant. Because of differences in group size and response rate, item difficulties for each group were anchored based on the item difficulties generated by the set of participants who responded to every item ($n = 37$). The characteristics of the students on whom the test were anchored were a reasonably close reflection of the larger sample. Specifically, there were 21 males, 15 females, and one student who failed to indicate their sex. In terms of ethnicity, 14 students were White, 14 were Asian–American or Pacific Islanders, one was Latino, two were African–American, and five students did not respond.

Although a variety of techniques are available for imputing missing data (e.g., Little & Rubin, 1987), there are many potential problems with these techniques, particularly if one suspects that the data are not missing completely at random. Consequently, because we were interested in estimating student ability, we chose to run a Rasch analysis for those students with complete data, anchor the item difficulty values, and then re-estimate person ability for all students in the sample based on the anchored item difficulty values. This approach is defensible because the Rasch technique is a sample-free approach to measurement meaning that so long as participants are drawn from the same population, the order of the item difficulty is presumed to remain invariant (see Bond & Fox, 2001; Wright & Stone, 1979). Furthermore, past research has shown that relatively stable item-difficulty estimates (95% CI ± 1 logit) can be obtained with approximately 30 participants (Linacre, 1994). Thus, we believe that this approach is somewhat stronger than imputation because it relies only on the data that students actually provide. That is, the student ability estimate is based only on the items each student attempted to answer and the person ability estimate derived is relative to the difficulty level of the items answered correctly. After individual person ability estimates were calculated for the entire sample, ethnic group differences in ability were then analyzed. An ability estimate of 0.00 logits indicates that the level of ability was exactly matched to the level of item difficulty; a negative estimate indicates that person ability was lower than the item difficulty (or the items were harder than the test-takers were capable of answering correctly); a higher estimate indicates that the person ability was higher than the item difficulty (or items were easier than those the test-takers could have

Table 4
Descriptive statistics of abilities by ethnic group.

	Range	Mean	*SD*
White students (N = 87)			
Whole test	2.52	0.31	0.40
Memory subscale	5.71	0.54	1.50
Analytic subscale	5.37	−0.68	0.80
Creative subscale	5.71	−0.31	0.97
Practical subscale	2.00	−0.22	0.41
African–American students (N = 32)			
Whole test	1.84	−0.73	0.41
Memory subscale	4.35	−0.67	1.27
Analytic subscale	3.18	−1.07	0.64
Creative subscale	4.20	−0.85	1.03
Practical subscale	1.84	−0.68	0.61
Latino students (N = 40)			
Whole test	2.18	−0.55	0.38
Memory subscale	5.17	−0.45	1.22
Analytic subscale	5.35	−1.02	0.93
Creative subscale	4.45	−0.72	1.11
Practical subscale	3.27	−0.52	0.57
Asian–American students (N = 69)			
Whole test	2.55	−0.42	0.50
Memory subscale	5.71	0.36	1.48
Analytic subscale	7.42	−0.67	0.95
Creative subscale	4.86	−0.49	1.13
Practical subscale	3.83	−0.42	0.62

Table 5
Differences in ability based on ethnic group.

	t-Statistic	Significance	Effect-size (Cohen's *d*)
African–American students v. White students			
Whole test	−5.08	.00**	−1.05
Memory subscale	−4.06	.00**	−0.87
Analytic subscale	−2.51	.01**	−0.54
Creative subscale	−2.68	.01**	−0.55
Practical subscale	−4.80	.00**	−0.90
Latino students v. White students			
Whole test	−3.13	.00**	−0.61
Memory subscale	−3.64	.00**	−0.72
Analytic subscale	−2.13	.04*	−0.40
Creative subscale	−2.12	.04*	−0.39
Practical subscale	−3.45	.00**	−0.62
Asian–American students v. White students			
Whole test	−1.51	.13	N/A
Memory subscale	−0.72	.47	N/A
Analytic subscale	0.04	.97	N/A
Creative subscale	−1.11	.27	N/A
Practical subscale	−2.47	.02*	−0.40

Note: Cohen's *d* statistics are computed using White students as the reference group.
Sample sizes for each ethnic group are: White students = 87, African American Students = 32, Latino Students = 40, and Asian students = 69.
* Difference is significant at the 0.05 level (2-tailed).
** Difference is significant at the 0.01 level (2-tailed).

answered). The range of ability estimates as well as the means and standard deviations by ethnic group appear below (Tables 4 and 5), followed by the differences in achievement across ethnic groups.

White students significantly out-performed African–American students on each subsection of the test as well as on the test as a whole. Although the effect-size difference between African–American students and White students was large on the test as a whole and on the memory and practical subscales, it was greatly reduced on the analytical and creative subscales. White students significantly outperformed Latino students on the test as a whole and on each subsection of the test. The sizes of these differences were moderate on the analytical, creative, and practical subsections, but noticeably larger on the memory subsection. White and Asian–American students did not perform significantly differently on the test as a whole. The only subscale on which White and Asian–American students showed a reliable difference in measured ability was on the practical subscale, where Asian–American students performed worse than White students by a relatively large amount. What these results seem to suggest is that an emphasis on memory-based items tends to lead to greater ethnic differences in student achievement than are observed when a broad range of cognitive skills are assessed.

3.2.3.2. Cognitive profiles of achievement. 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-type 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. Mathematically, the technique is identical to traditional factor analysis (also known as R-type factor analysis). The main difference is that with the Q-type approach individuals are factored together in the way that items typically would be in an R-type approach. Using principal-components analysis with a promax rotation, we obtained three factors with eigenvalues greater than 1.0 that accounted for 100% of the variance in the data. Promax rotation was used because we expected that the profiles (i.e., factors) extracted would be not be completely orthogonal. The three factors correspond to three distinct profiles of achievement represented in the dataset. Table 6 presents an abridged structure matrix of factor loadings for each participant on the exam. Looking at the table, we can see that some people had high positive loadings on a single factor whereas other students had high factor loadings in the negative direction.

To demonstrate the meaning of 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 in the second row, on the second factor; and the four participants in the third row, on the third factor.

As can be seen by examining Fig. 3, those participants with high loadings on the first factor tended to exhibit relatively high performance on the memory subscale, with lower scores on the analytical, creative, and practical subscales. Their performance on the analytical subscale was slightly better than their performance on the creative or practical subscales. They thus seem to fit a more traditional pattern of abilities, with emphasis on memory and analytical skills (Sternberg, 1997).

Participants with high loadings on the second factor demonstrated high performance on the memory and creative subscales, lower performance on the practical subscale, and very low performance on the analytical subscale. These students thus tended toward a more creative profile (Sternberg, 1997).

Finally, participants with high loadings on the third factor exhibited high scores on the practical subscale, low scores on the creative subscale, and roughly equivalent scores on the memory and analytical subscales. These students thus tended toward a more practical profile (Sternberg, 1997).

The patterns described above are consistent with the theory of successful intelligence (Sternberg, 1997) and suggest that the three abilities posited by the theory––analytical, creative, practical––which previously were extracted by R-factor analysis (e.g., Sternberg & The Rainbow Project Collaborators, 2006), also emerge through Q-factor analysis.

Table 6
Abridged output of factor loadings for each participant.

Student	Component		
	1	2	3
K_232	1.000		
K_112	1.000		
K_179	.999		
K_37	−.997		
K_221	.996		
K_114	.996		
K_3	.995		
K_72	.994		
K_249	.992		
K_235	.992		
K_169		.984	
K_51		−.984	
K_271		.983	
K_280		.982	
K_281		.980	
K_45		.980	
K_58		.976	
K_276		.976	
K_181		.976	
K_34		.975	
K_129			.992
K_209			.988
K_7			.987
K_189			−.987
K_149			−.976
K_86			.961
K_141			−.960
K_89			.934
K_126			.931
K_46			.926

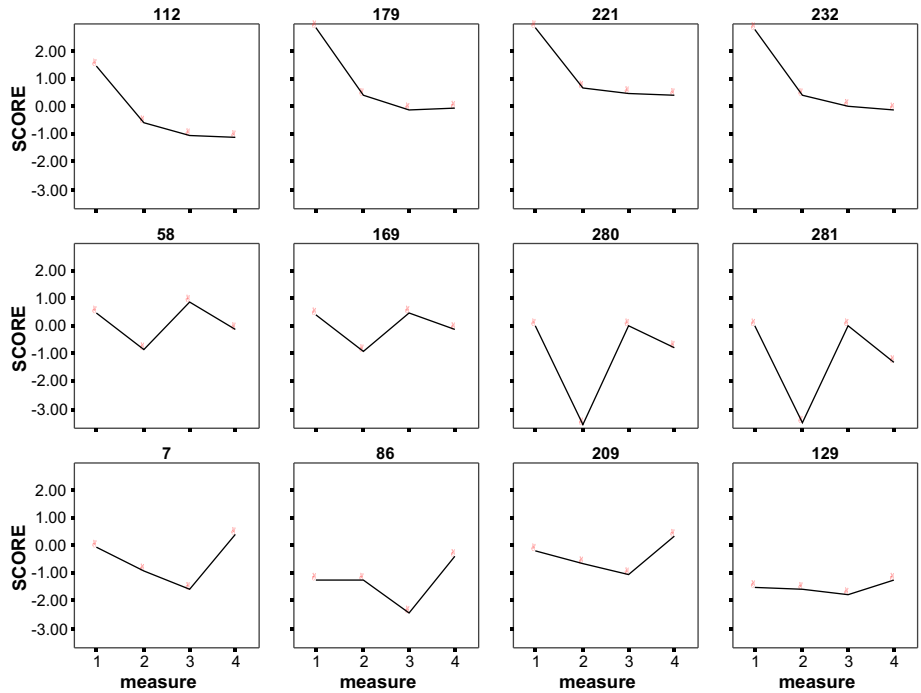

Fig. 3. Exemplary profiles of achievement. *Note*. On the *X*-axis for each figure, 1, memory subscale score; 2, analytic subscale score; 3, creative subscale score; 4, practical subscale score. The unit of measure on the *Y*-axis is logits.

It is important to note that, whereas some participants had high positive loadings on a factor, others had negative loadings on the same factor. That is, participants with negative loadings on a factor demonstrated the opposite pattern of achievement as those participants with positive loadings on the factor. For example, participants with high positive loadings on the third factor demonstrate high achievement on the practical subscale and low achievement on the creative subscale, whereas participants with high negative loadings show high achievement on the creative subscale and low achievement on the practical subscale. Therefore, although three factors were extracted, they yielded six distinct profiles of achievement.

Table 7 presents a summary of the number of participants whose profiles were associated with each of the six empirically distinct profiles of achievement. The most common profile type is associated with strong memory skills and somewhat strong analytical skills (Profile 1––positive loading), but many participants exhibited profiles that were not characterized by strong memory or analytical skills.

A chi-square test of association revealed that the actual number of participants associated with each profile divided by sex and ethnicity was not significantly different from the expected number of participants, $\chi^2(15, N = 228) = 8.77$, $p = .89$ for ethnicity and $\chi^2(5, N = 256) = 1.36$, $p = .93$ for sex. Therefore, these different profiles were not systematically associated with the individual characteristics of sex or ethnicity. Tables 8 and 9 summarize the number of participants associated with a particular profile of achievement, broken down by ethnicity and sex, respectively.

Table 7
Summary of participants associated with each profile.

Profile	Frequency	Percent (%)
Profile 1 – *HM*	107	38.1
Profile 2 – *HAHPHC*	41	14.6
Profile 3 – *HMHC*	61	21.7
Profile 4 – *HAHP*	9	3.2
Profile 5 – *HMHP*	40	14.2
Profile 6 – *HC*	23	8.2

Note: HM, high memory; HAHPHC, high analytic, high creative, high practical; HMHC, high memory, high creative; HAHP, high analytic, high practical; HMHP, high memory, high practical; HC, high creative.

4. General discussion

Based on classical test theory and the Rasch model, it appears that the items on the augmented version of the AP Physics examination are functioning well. Both sets of statistics indicate that, in general, the people who were predicted to answer an item correctly were doing so, and that those who were not, were answering incorrectly. Additionally, the data demonstrate that the items vary substantially in difficulty levels, although both classical-test-theory statistics and the results of the Rasch analysis suggest that there are too many difficult items on the exam, particularly with regard to the ability level of the test-takers. This is of particular importance for the Rasch model, which stipulates that the items closest to a person's ability level are the most effective at estimating person ability. It is possible, however, that this particular version of the test would function even more effectively if administered to the broad range of AP Physics students (exhibiting wider variability in achievement) nationwide. In addition, the goal of the exam must always be kept in mind. The AP Physics exam is meant to identify students with exceptional knowledge of the subject matter; thus, having more difficult items than might normally be ideal is in line with the goal of the assessment. Nevertheless, a limitation of criterion-referenced testing in general is that when the criterion is set higher than the ability level of most students taking the test, the observed reliability of the instrument will be somewhat attenuated. Furthermore, as the idea of testing subject matter expertise from multiple cognitive perspectives is relatively new, this may well have been the first time many of the test-takers were exposed to test items that required them to use primarily creative or practical thinking processes. Although the high difficulty of the assessment relative to the ability of the test-takers is a

Table 8
Summary of participants associated with each profile by ethnicity.

Profile		Ethnicity				
		African–American	White	Latino	Asian–American	Total
Profile 1 – *HM*	Count	10	41	12	25	88
	% of total	4.4	18.0	5.3	11.0	38.6
Profile 2 – *HAHPHC*	Count	8	13	6	11	38
	% of total	3.5	5.7	2.6	4.8	16.7
Profile 3 – *HMHC*	Count	7	16	9	12	44
	% of total	3.1	7.0	3.9	5.3	19.3
Profile 4 – *HAHP*	Count	1	2	1	2	6
	% of total	0.4	0.9	0.4	0.9	2.6
Profile 5 – *HMHP*	Count	5	10	9	13	37
	% of total	2.2	4.4	3.9	5.7	16.2
Profile 6 – *HC*	Count	1	5	3	6	15
	% of total	0.4	2.2	1.3	2.6	6.6
Total	Count	32	87	40	69	228
	% of total	14.0	38.2	17.5	30.3	100.0

Note: HM = high memory; HAHPHC = high analytic, high creative, high practical; HMHC = high memory, high creative; HAHP = high analytic, high practical; HMHP = high memory, high practical; HC = high creative.

Table 9
Summary of participants associated with each profile by sex.

Profile		Sex		
		Male	Female	Total
Profile 1 – *HM*	Count	61	42	103
	% of total	23.8	16.4	40.2
Profile 2 – *HAHPHC*	Count	27	14	41
	% of total	10.5	5.5	16.0
Profile 3 – *HMHC*	Count	28	20	48
	% of total	10.9	7.8	18.8
Profile 4 – *HAHP*	Count	4	3	7
	% of total	1.6	1.2	2.7
Profile 5 – *HMHP*	Count	21	17	38
	% of total	8.2	6.6	14.8
Profile 6 – *HC*	Count	10	9	19
	% of total	3.9	3.5	7.4
Total	Count	151	105	256
	% of total	59.0	41.0	100.0

Note: HM, high memory; HAHPHC, high analytic, high creative, high practical; HMHC, high memory, high creative; HAHP, high analytic, high practical; HMHP, high memory, high practical; HC, high creative.

limitation of this study, until the field finds alternative means of establishing the reliability of challenging measures presented to those who are not yet competent, we will confront this problem quite routinely.

In terms of overall test functioning, the two sets of statistics suggest that this assessment is reliable in terms of its potential for consistently measuring person ability and item difficulty. The reasonably high level of internal-consistency reliability of the overall test suggests that the items are holding together to measure the latent trait of physics knowledge, and the item analysis results suggest that each item is importantly contributing to this measurement. Additionally, the results of Study 1 support the content validity of the newly developed items on the augmented AP Physics examination in that the newly developed items on the exam were rated by AP Physics teachers as tapping important content to the domain and were rated by cognitive experts as adequately tapping the broad range of cognitive-processing skills that include memory, analytical, creative, and practical skills. Although there is some variation between raters with regard to their use of the rating scale, on the whole, the raters tended to exhibit acceptably high levels of consensus interrater reliability (e.g., percent agreement values generally in the .70s and positive kappa values). That both subject matter and process experts reliably categorized the items as belonging to each of the four cognitive process subscales suggests that the test had good content validity (i.e., experts agree that the test measured what it claimed to measure). Overall, the findings of both Studies 1 and 2 suggest that the augmented version of the AP Physics exam has good internal psychometric properties.

It appears as though balancing tests for both content and cognitive-processing skills yielded benefits at both the individual and group levels. At the individual level, this study revealed that students exhibited six distinct profiles of achievement in AP Physics that corresponded to different patterns of high and low achievement across the different processes. This is fully consistent with the findings of Stemler et al. (2006), who identified six profiles of achievement among students on AP exams in psychology and statistics and were consistent with the theory of successful intelligence (Sternberg, 1997). More remarkably, the profiles exhibited by students on AP psychology, statistics, and physics exams were exactly the same. This result underscores the importance of developing assessments that measure a broad range of cognitive processes, for two reasons. First, that students exhibit consistent strengths and weaknesses across diverse subject matters provides compelling evidence that cognitive-process skills can be considered relatively independent of content. Thus, assessments that measure content alone are incomplete indicators of students' knowledge; it is inappropriate to test content without considering the impact of cognitive processes. Second, although the most dominant profile of student achievement was characterized by relatively strong memory and analytical skills, this profile represented only 38% of students. Therefore, many students exhibited profiles that were not associated with strong memory or analytical skills; traditional tests that focus largely or even exclusively on these two process areas may fail to detect the relative strengths of individuals showing other profiles of skills.

As expected, the mean ability levels on the subscales for each profile group were relatively higher on the subscales that represented processes where groups were comparatively strong and lower on those where groups were comparatively weak. So, although overall students found certain subscales to be easier than others (in particular, students found the memory subscale to be the easiest followed by the practical, creative, and analytical subscales), this appears to be an artifact of the proportion of students

characterized by certain profile types rather than evidence of a hierarchical progression of cognitive skills. Furthermore, there appears to have been no discernible relationship between profile type and score on the overall assessment, given that the two highest scoring profile groups and the lowest scoring group had similar strengths and weaknesses. Specifically, the two highest scoring groups were characterized by strong memory and creative skills, respectively, but the lowest scoring group exhibited relative strengths in both memory and creative skills. It may not be surprising that there was no relationship between strengths and weaknesses in particular cognitive-processing skills and overall score on this test, given that it was intended to measure a broad range of cognitive-processing skills, through which it was expected that students with a variety of profile types would be allowed to demonstrate their content mastery. As the present study lacked information regarding students' performance on the actual AP Physics exam, it is not possible to say whether overall score on the actual assessment would differ based on profile type and, thus, no direct comparison between the tests can be made.

The observed profiles were independent of sex and ethnicity, so students across demographic groups exhibited the same patterns of strengths and weaknesses in approximately the same proportions. This finding provides further support for the idea that these profiles are broadly generalizable; in addition to being unrelated to content, they appear to be relatively invariant across demographic groups. Across ethnic groups, people find the same cognitive processes relatively harder or easier; but a number of ethnic differences were observed in performance on the overall test and the subscales. In particular, White students out-performed African–American and Latino students on all subscales and Asian–American students on the practical subscale. Although these results initially appear discouraging, examination of the effect-sizes suggests that augmenting the AP Physics exam had an important impact on reducing the typically observed achievement gap. Recall that the effect-size difference between African–American students and White students is approximately one standard deviation on most traditional tests of achievement. Performance on the current test seems to replicate these findings: The effect-size difference on the overall test, and the memory and practical subscales were roughly one standard deviation. However, the effect-size difference on the analytical and creative subscales was only half that of standard estimates. Thus, the discrepancy in student achievement in White and African–American students that is typically observed on exams that stress analytical and memory skills would be only about half as much after the introduction of a creative subscale. This result is consistent with the findings of Stemler et al. (2006), who found that differences between African–American and White students on the AP Psychology and AP Statistics exams were significantly reduced on the creative subscale, and with the findings of the Sternberg and The Rainbow Project Collaborators (2006) study demonstrating reduced differences between African–American and White students on tests of creativity in the context of college admissions. Thus, it appears as though including a creative subscale benefits African–American students, regardless of content domain (at least for the content domains we have studied).

One surprising finding in light of past data (Jencks & Phillips, 1998) was the reliably low test-score difference between White students and African–American students on the analytical-reasoning subscale. One possible explanation for this unexpected finding relates to selection bias. Recall that one of the important challenges facing the AP program is the enrollment of minority students; only a very small percentage of African–American students enroll in the AP program in general, and in AP Physics in particular, compared with their representation in the overall student population. Given that most students are likely selected into AP Physics on the basis of their strong analytical skills (as demonstrated through high performance on past exams), the analytical skills of African–American students selected into the AP program may be extremely high within their own ethnic group. By contrast, given the greater proportion of White students enrolled in AP Physics, it is possible that White students may represent a broader range of the spectrum for their ethnic group so that the small ethnic differences observed on the analytical section may be an artifact of pre-existing selection differences. This interpretation is merely speculative, however, and should be investigated in future research.

Latino students exhibited moderate differences on the analytical, creative, and practical subsections compared with White students, but noticeably larger differences on the memory subsection. This finding is also consistent with previous research on the AP Psychology and AP Statistics exams, where the largest difference between Latino students and White students was observed on the memory subscales of the augmented AP exams (Stemler et al., 2006). The typically observed achievement gap between Latino students and White students ($d = -0.58$) is somewhat reduced on both the analytical and creative subscales, although not on the practical subscale. This result represents a small change in measured ability, but most people would agree that any reduction in the achievement gap (even a small one) is beneficial. Interestingly, Asian–American students performed moderately, though significantly, worse than White students on the practical subscale.

Overall, it appears that augmenting the AP Physics exam with creative and practical subscales effectively allowed underrepresented minorities to better express their content knowledge.

There a number of noteworthy limitations to the present research. First, without rater agreement results for more subject-matter experts, this study leaves open the question of whether agreement for subject-matter experts one and two or agreement for subject matter expert three with either of the other two experts is more typical. In addition, data on students' actual AP Physics exam scores or college outcomes were not available. As a result, this study was not able to discuss the relationship between the actual AP exam and this augmented version. Furthermore, actual AP exam data would allow a stronger argument to be made regarding the reduction in achievement differences across ethnic groups that resulted from using the augmented version of the AP Physics exam.

5. Conclusions

This research demonstrates the usefulness and importance of developing tests that measure a broad range of cognitive skills on a number of fronts. First, the research provided some evidence to suggest that AP Physics teachers can be trained reliably to develop items that are specifically balanced not only for content but also for cognitive-processing skills demanded to correctly answer an item. Both independent subject-matter experts and cognitive experts stated with greater than 70% agreement that the newly developed items measure information and processing skills that AP Physics teachers feel is worth knowing. Next, ethnic differences in achievement were significantly reduced on certain subscales. As scores on the AP exam have real effects on college admission, performance, and course choice (Dodd et al., 2002; Dougherty, Mellor, & Jian, 2005; Gieiser & Santelices, 2004; Morgan & Maneckshana, 2000; Morgan & Ramist, 1998), improved performance in underrepresented minority groups could dramatically alter the ethnic make-up of academic departments and colleges (and, eventually, professional fields). There is a growing body of research on the positive externalities of diverse educational environments (e.g., Shaw, 2005), so a reduced achievement gap not only stands to benefit underrepresented students but their classmates and society at large. Third, the existence of distinguishable profiles of achievement demonstrates that students have consistent patterns of

strengths and weaknesses across cognitive-process areas that are independent of content, and that are consistent with the theory of successful intelligence (Sternberg, 1997). Thus, in order to make more valid inferences about students' content mastery, tests need to take these cognitive-processing skills into account. Because many of these profiles are not characterized by strong skills in the cognitive-process areas that are emphasized by traditional assessments (i.e., memory or analytical skills), unless measuring a range of skills is an explicit goal in test development, large numbers of students will not be permitted to fully demonstrate their mastery of a subject area. Overall, this study not only suggests that it is possible to ground an AP Physics exam in a modern theory of cognitive processing, but that doing so yields many noteworthy benefits.

The present study expands on previous research that has consistently demonstrated the many benefits of using cognitive-based assessments in the classroom (Sternberg & Grigorenko, 2000), on the SAT (Sternberg & The Rainbow Project Collaborators, 2006), and on AP exams (Stemler et al., 2006). The results provide hope for those who wish to maximize educational outcomes while at the same time optimizing equity in assessment.

Acknowledgments

The work reflected in this paper and the paper's preparation was supported by grants REC-0440171 (Yale University) and REC-0710915 (Tufts University) from the National Science Foundation. Grantees undertaking such projects are encouraged to express freely their professional judgment. This paper, therefore, does not necessarily represent the position or policies of the National Science Foundation, and no official endorsement should be inferred. We are grateful to all the students, teachers and school administrators for their participation in this project. We also thank Carolyn Parish, Derrick Carpenter, Arrash Baghaie, Kathleen Connolly, and Gary Stemler for their contributions to this research.

References

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.

Carraher, T. N., Carraher, D., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3*, 21–29.

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

Clark, D. (1999, May 21, 2000). *Bloom's taxonomy*. <http://www.nwlink.com/~donclark/hrd/bloom.html> Retrieved 27.04.04.

College Board. (2004). *Exam scoring*. <http://apcentral.collegeboard.com/article/0,3045,152-167-0-1994,00.html> Retrieved 20.05.04.

College Board. (2007). *Advanced placement report to the nation*. <http://apcentral.collegeboard.com> Retrieved 15.12.07.

Dodd, B. G., Fitzpatrick, S. J., DeAyala, R. J., & Jennings, J. A. (2002). *An investigation of the validity of AP grades of 3 and a comparison of AP and non-AP graduate groups*. New York: College Entrance Examination Board.

Dougherty, C., Mellor, L., & Jian, S. (2005). *The relationship between Advanced Placement and college graduation (Report No. 1)*. Austin, TX: National Center for Educational Accountability.

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

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

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. O., Mullis, I. V. S., et al. (2006). *TIMSS advanced 2008 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Gieiser, S., & Santelices, V. (2004). *The role of advanced placement and honors courses in college admissions*. Berkeley, CA: Center for Studies in Higher Education.

Gronlund, N. (2002). *Assessment of student achievement* (7th ed.). New York: Pearson Education.

Guyote, M. J., & Sternberg, R. J. (1981). A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology, 13*, 461–525.

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

Hanushek, E. A., & Rivkin, S. G. (2006). *School quality and the Black–White achievement gap* [Electronic Version]. MBER working paper 12651. <http://www.nber.org/papers/w12651>.

Hedges, L. V., & Nowell, A. (1998). Black–White test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 149–181). Washington, DC: Brookings Institute Press.

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

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

Klopfenstein, K. (2004). Advanced placement: Do minorities have equal opportunity? *Economics of Education Review, 23*, 115–131.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The nature of reasoning*. New York: Cambridge University Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Martin, M. O., Mullis, I. V. S., & Foy, P., (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mislevy, R. J. (2004). *The case for an integrated design framework for assessing science inquiry (Report No. CSE 638)*. Los Angeles, CA: Center for the Study of Evaluation (CSE); National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Morgan, R., & Maneckshana, B. (2000). *AP students in college: An investigation of their course-taking patterns and college majors (Report No. 2000-09)*. Princeton, NJ: Educational Testing Service.

Morgan, R., & Ramist, L. (1998). *Advanced placement students in college: An investigation of course grades at 21 colleges. (Report No. 98-13)*. Princeton, NJ: Educational Testing Service.

Nettles, A. L., & Nettles, M. T. (Eds.). (1999). *Measuring up: Challenges minorities face in educational assessment*. Boston: Kluwer Academic.

Mullis, I. V. S., Martin, M. O., & Foy, P., (with Olson, J. F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Nuñes, T. (1994). Street intelligence. In R. J. Sternberg (Ed.). *Encyclopedia of human intelligence* (Vol. 2, pp. 1045–1049). New York: Macmillan.

Nuñes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. New York: Cambridge University Press.

Perret-Clermont, A. N. (1980). *Social interaction and cognitive development in children*. London: Academic Press.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic group. College Board Report No. 93-1*. New York: College Board.

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

Shaw, E. (2005). *Researching the educational benefits of diversity (No. 2005-4)*. New York: College Entrance Examination Board.

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. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). <http://pareonline.net/getvn.asp?v=9&n=4>.

Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in psychology and statistics. *Contemporary Educational Psychology, 31*(2), 75–108.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks, CA: Sage.

Sternberg, R. J. (1980). Intelligence and test bias: Art and science. *Behavioral and Brain Sciences, 3*, 353–354.

Sternberg, R. J. (1982). Nonentrenchment in the assessment of intellectual giftedness. *Gifted Child Quarterly, 26*, 63–67.

Sternberg, R. J. (1983). Components of human intelligence. *Cognition, 15*, 1–48.

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., & Ben-Zeev, T. (2001). *Complex cognition: The psychology of human thought*. New York: Oxford University Press.

Sternberg, R. J., & The Rainbow Project Collaborators (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, creative, and practical skills. *Intelligence, 34*(4), 321–350.

Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg triarchic abilities test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment, 17*(1), 1–16.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., Wagner, R. K., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General, 112*, 80–116.

Sternberg, R. J., & Gastel, J. (1989a). Coping with novelty in human intelligence: An empirical investigation. *Intelligence, 13*, 187–197.

Sternberg, R. J., & Gastel, J. (1989b). If dancers ate their shoes: Inductive reasoning with factual and counterfactual premises. *Memory and Cognition, 17*, 1–10.

Sternberg, R. J., & Grigorenko, E. L. (2000). *Teaching for successful intelligence*. Arlington Heights, IL: Skylight Professional Publishers.

Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. (1999). A triarchic analysis of an aptitude–treatment interaction. *European Journal of Psychological Assessment, 15*, 1–11.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York: Free Press.

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.

Sternberg, R. J., Wagner, R. K., & Okagaki, L. (1993). Practical intelligence: The nature and role of tacit knowledge in work and at school. In H. Reese & J. Puckett (Eds.), *Advances in lifespan development* (pp. 205–227). Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912–927.

Uebersax, J. (2002). *Statistical methods for rater agreement*. <http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm> Retrieved 9.08.02.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality & Social Psychology, 52*, 1236–1247.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436–458.

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

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