

Using modern theories of cognitive processing to
augment assessments in Advanced Placement Physics

by

Kirsten Sharpes
Class of 2008

A thesis submitted to the
faculty of Wesleyan University
in partial fulfillment of the requirements for the
Degree of Bachelor of Arts
with Departmental Honors in Psychology

Acknowledgements

I am fortunate to have received a great deal of support from professors, peers, family, and friends throughout my time at Wesleyan. There are a few people who deserve special mention for the exceptional contributions that they have made to this thesis and to my academic career.

For his extensive assistance at all stages of the research process, I sincerely thank Professor Steve Stemler. He devoted an extraordinary amount of time and energy to guiding me in my research and helping me develop analytical tools that I am sure will serve me well for years to come. This has been an incredibly rewarding and inspiring research relationship.

I have also benefited greatly from working with Professor Andrea Patalano. She has provided me with tremendous support and encouragement throughout this process. I am grateful for the having had the opportunity to be a part of her lab for so many years and have learned so much from working with her.

Manolis Kaparakis deserves thanks for his warm-hearted support and good advice. His encouragement to pursue quantitative research has had a huge impact on my academic career.

I have had the opportunity to work with a number of remarkable students during my time at Wesleyan who taught me a great deal. In particular, I thank Steven Wengrovitz for his friendship, guidance, and amazing editorial skills. He is the reason I became involved in psychology research, for which I am incredibly grateful.

Finally, I could never express how much I appreciate the love and support of my family. I especially wish to thank my husband, Dusty Sharpes, for his never-ending patience, encouragement, and devotion, which were instrumental in enabling me to complete this project.

Abstract

The current research examined whether theory-driven examinations that expand the range of cognitive skills assessed allow individual students to better demonstrate content knowledge and decrease achievement gaps between ethnic and sex groups. Two hundred and eighty one students enrolled in AP Physics courses in the 2006-2007 school-year took an augmented version of the AP Physics exam that included items measuring creative and practical thinking in addition to analytic and memory skills. Employing such a framework reduced achievement differences in ethnic groups compared to standard estimates. It further revealed distinct profiles of achievement across cognitive processes, suggesting that traditional tests, which tend to emphasize memory and analytic skills, may not allow all students to fully demonstrate their content mastery. This research points to a need to integrate theories of cognitive processing into the design of new tests.

Using modern theories of cognitive processing to augment assessments in Advanced
Placement Physics

Although achievement testing can be traced back to about 150 BC as a means of selecting government employees in China (Bowman, 1989), the testing boom that is so evident today got its start in the early 20th century. E.L. Thorndike is often considered the father of the educational testing movement as he was involved in the development of many of the first standardized achievement tests (e.g., Thorndike's Scale of Handwriting for Children) when previously testing had been highly subjective (Ross & Stanley, 1954). Over the course of the century, the use of objective, standardized testing increased as researchers became more aware of the poor reliability of teachers' assessment (e.g., Starch & Elliot, 1913 as cited in Wiliam, 2006). Indeed, in 1926 the Scholastic Aptitude Test (SAT) replaced the essay tests that had previously been required of college applicants (Donlon, 1984). The prominence of standardized testing continued to rise over the rest of the century: the Advanced Placement (AP) program was introduced in 1955 (College Board, 2008a) and the ACT, an alternative to the SAT, was introduced in 1959 (ACT, 2008). These tests represent some of the largest, richest, and most consequential in the field of educational testing.

Today standardized tests are frequently used in the United States and abroad as a basis for making decisions about the educational opportunities, placement, and diagnosis of students. 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). Much as the birth of the modern

standardized testing movement came about because of the shortcomings of subjective assessment, standardized achievement tests are frequently used to make important decisions (e.g., college admissions or school funding) when other criteria are more subjective or less easily comparable (e.g., teacher evaluations or high school GPA). In the context of college admissions, researchers, though careful not to diminish the value of subjective indicators, stress the usefulness of standardized tests in adding to predictions of college success (Bejar & Blew, 1981; Bridgeman, Pollack, & Burton, 2004; Camara & Echternacht, 2000).

One prominent player in the standardized testing arena is the College Board's Advanced Placement (AP) program. This program began in 1955 as a way to expose advanced high school students to college level material, grant them college credit for their work, and allow them to bypass introductory level courses in college (College Board, 2008a). 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, 2008a).¹

Each spring, students enrolled in AP courses are given the opportunity to take a high-stakes examination to demonstrate their mastery of a subject area. The AP exams are graded on a 1 to 5 scale, with score of 5 indicating “a student who is

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, U.S. 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, U.S. History, and World History.

extremely well-qualified to receive college credit and/or advanced placement based on an AP exam grade” (College Board, 2004). More than 90% of four-year colleges and universities in the United States allow students scoring 3 or higher on an exam to receive college credit, bypass an introductory level course, or both in that subject area (College Board, 2008a). Thus, the results of the test have important financial implications, as placing out of introductory college courses could save a student thousands of dollars in tuition in subsequent years. Furthermore, AP scores are frequently used in admissions decisions as evidence of commitment to academic excellence and as predictors of success in college. Recent studies suggest that students who score well enough on AP exams to place out of introductory courses are more likely to graduate college in five years or less, pursue higher-level courses in the exam subject, and succeed academically in college (Dodd, Fitzpatrick, DeAyala, & Jennings, 2002; Dougherty, Mellor, & Jian, 2005; Geiser & Santelices, 2004; Morgan & Maneckshana, 2000; 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.

Although AP tests have many benefits associated with standardization, the program still has some noteworthy weaknesses, two of which are the focus of the present research. A discussion of the importance of grounding assessments in established theories of cognitive processing, as well as a review of the literature on

bias in standardized testing, is presented below prior to a description of how the present research addresses these issues.

The lack of theoretical basis in testing

Despite the large number of students who take standardized achievement tests every year, many of these tests are not aligned with modern theories of learning and cognitive processing. These tests tend to emphasize only a limited range of skills (e.g., analytical and memory skills) and, as a result, students with strengths in cognitive processes that are not measured by these tests may not be able to fully express their content knowledge (Sternberg, 1997). Specifically, many students with strong creative or practical skills are less able to demonstrate these abilities on conventional examinations, despite the importance of diverse skill sets. This situation is especially problematic because 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, it takes creative skills to synthesize disparate findings and devise new theories, and practical skills to understand how theoretical findings may be applied to solve real world problems. Indeed, a balance of cognitive skills is important, regardless of one's professional domain and, accordingly, a broad range of skills should be assessed in students.

Historically, a chief concern of AP exam developers has been 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). Feedback to test-takers might indicate the specific content areas where they are strong (e.g., electricity) or weak (e.g., thermodynamics). Traditionally, however, there has been no systematic attempt to explicitly balance items for the cognitive processing skills they assess. Only in recent years have designers of large-scale testing programs become interested in linking educational assessment to modern theories of cognitive processing (Embretson & Reise, 2000; Irvine & Kyllonen, 2002).

In today's high-stakes testing environment where there are important social, economic, and ethical consequences associated with standardized testing, a test lacking an underlying cognitive theory may fail to adequately measure the content mastery of all students and, thus, may unduly curtail their future opportunities. This is particularly the case as large-scale assessments are increasingly expected to serve as tools for diagnosing students' cognitive strengths and weaknesses. For example, under the No Child Left Behind Act of 2001, teachers and administrators are expected to use the results of the annual assessments to "diagnose and meet the needs of each student" (United States Department of Education, 2004, 3rd paragraph). In addition, leading educational organizations such as the National Research Council have recognized the importance of grounding assessments in cognitive theories of learning (see Leighton & Gierl, 2007). Unfortunately, few large-scale tests are ready to fulfill their expected roles as providers of cognitive diagnostic information because few are developed with this purpose in mind, let alone with a theoretical grounding in construct-validated theories of cognitive processing (Lane, 2004; Leighton, 2004). These concerns apply equally to the AP program (Gollub, Bertenthal, Labov, &

Curtis, 2002; Labov, 2002). The current version of the AP program focuses primarily on expert-selected content rather than the thinking processes that are concurrently measured by the examination. But it is a mistake to assume that content knowledge alone is responsible for a student's ability to solve a problem on the AP exam; surely the type of thinking required has a critically important effect as well. An AP physics student with strong practical skills and weak analytic skills might demonstrate great mastery of thermodynamics on applied problems but perform poorly on questions requiring analytical thinking in the same content area. The misguided view that only content matters also affects the material covered in AP classrooms: AP teachers are not instructed to think about learning principles in teaching, and thus may not be meeting the needs of all students as well as they could be.

Tests that have made an attempt to explicitly measure students' cognitive skills (Beaton et al., 1997; Mullis et al., 1997) have traditionally used Bloom's *taxonomy of the cognitive domain* (Bloom, 1956) as a theoretical foundation. Within this framework, students' intellectual skills are thought to progress hierarchically, beginning with knowledge of a topic and proceeding through the stages of comprehension, application, analysis, synthesis, and evaluation. *Knowledge* is demonstrated through the recall of information. *Comprehension* is demonstrated when an individual restates a problem in his own words. *Application* is demonstrated when an individual applies what he has learned to a novel situation. *Analysis* is demonstrated when an individual separates a problem into component parts or identifies an underlying structure to a problem. *Synthesis* occurs when an individual re-combines separate parts to create a new whole. *Evaluation* is demonstrated when

an individual makes judgments about the value of ideas (Clark, 1999). Thus, users of the taxonomy assume that if an individual is able to successfully analyze a problem, he will be able to apply his knowledge of the topic as well.

A more recent theory of cognitive processing is Sternberg's *theory of successful intelligence*. According to this theory, a common set of processes underlies all aspects of problem solving (Sternberg, 1984, 1985, 1997, 1999). *Metacomponents*, or executive processes, plan what to do, monitor activities as they are being completed, and evaluate outcomes after activities are complete. *Performance components* execute the instructions of the metacomponents. *Knowledge-acquisition components* are used to acquire declarative knowledge about how to solve problems. These processes are hypothesized to be universal although the problems and culturally acceptable approaches and solutions (i.e., whether a given problem requires analytical thinking, creative thinking, practical thinking, or a combination of these kinds of thinking) may vary by context. The types of thinking that may be employed are analytical, creative, and practical thinking. *Analytical* thinking is used when components are applied to fairly familiar kinds of problems abstracted from everyday life. On tests, students use analytical skills when they are asked to critique, evaluate, compare and contrast, or otherwise analyze; this is the primary cognitive process that traditional achievement tests such as the AP exam or SAT have assessed. *Creative* thinking is used when the components are applied to relatively novel kinds of tasks or situations. In the context of a test, creative items ask students to imagine, invent, discover, suppose, or create. *Practical* thinking is used when the components are applied to adapt to, shape, and select environments. Students use practical thinking on tests when they are asked to

apply, use, or otherwise put into practice. Thus, the same components, applied in different contexts, yield three different kinds of thinking: analytical, creative, and practical, with *memory* skills serving as a foundation for each type of thinking.

The two theories of cognitive processing are not necessarily incompatible. Indeed, the application level of Bloom's taxonomy is similar to the practical skills put forth in Sternberg's theory. In addition, the synthesis level of Bloom's taxonomy shares some features with the creative skills aspect of Sternberg's theory. One key distinction between the theories is that Bloom's taxonomy specifies a hierarchical progression of cognitive skills, whereas Sternberg's theory takes an interactive and profile-oriented approach. That is, the theory of successful intelligence suggests that it 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. As a result, for example, some students may be more capable of demonstrating their knowledge when problems are placed in a practical but not an analytical context. Others may show the reverse pattern. These implications have been borne out in applied research. For example, Nuñez and colleagues (Carragher, Carragher, & Schliemann, 1985; Nuñez, 1994; Nuñez, Schlieman, & Carragher, 1993) studied Brazilian children who, for economic reasons, often worked as street vendors and had little formal schooling. These children were successful in completing trade-related computations but were unable to solve computationally similar problems when they were presented in abstract terms. Conversely, many European schoolchildren were able to solve pencil-and-paper arithmetic questions but could not solve the same type of problem in an applied

context (Perret-Clermont, 1980). Similar results have been found in the literature on logical reasoning (Leighton & Sternberg, 2004; Sternberg & Ben-Zeev, 2001).

The theory of successful intelligence has been effectively used as a basis for enhancing classroom teaching (Sternberg & Grigorenko, 2000; Sternberg & Spear-Swerling, 1996), the SAT (Sternberg & The Rainbow Project Collaborators, 2006), and AP exams in psychology and statistics (Stemler, Grigorenko, Jarvin, & Sternberg, 2006). For example, Stemler and his colleagues (2006) found that the results of augmenting AP Psychology and Statistics exams with creative and practical subscales supported the construct validity of the theory of successful intelligence. In this study, the results of Q-type factor analyses revealed six distinct profiles of achievement among students, supporting the assertion that individuals exhibit different patterns of strengths and weaknesses across cognitive processing skills.

An advantage to using an expanded theory of cognitive-processing skills in test development is that this approach can provide useful information about individual students. Using this approach, test-takers could receive individual score reports that outline their profiles 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 correct for their weaknesses. Furthermore, by measuring a broader range of cognitive skills, students who might have been labeled as low-achievers when they were assessed based on a limited set of skills may have improved opportunities to demonstrate their mastery of a content area. To the extent that selection tests are weighted more heavily in favor of one type of skill, an entire professional field may suffer because it may be dominated by individuals with a

single profile of strengths and weaknesses, thereby inhibiting the capacity of the field to realize its full potential. Assessing a broader range of cognitive skills may also increase the diversity of thought and capacity for innovation in formerly more homogenous fields. In particular, prior research has shown that traditionally underrepresented minority students stand to benefit from broader measures of cognitive skills as detailed later (Sternberg & The Rainbow Project Collaborators, 2006; Sternberg, Torff, & Grigorenko, 1998).

Capitalizing on past research on the effectiveness of using the theory of successful intelligence as a basis for teaching and testing, the current project involves the development and analysis of an augmented test in AP Physics B that is explicitly linked to Sternberg's theory of successful intelligence.

Bias in testing

Arguably, one of the most prominent issues in testing is bias. Bias is a technical term that refers to deficiencies in a test that result in different meanings for scores earned by members of different identifiable groups (American Educational Research Association, 1999). Whereas colloquially bias may refer to differences in the average scores among ethnic, sex, or socioeconomic groups, bias from a psychometric perspective requires differential validity of scores. Differences in average scores alone do not mean that a test is biased, although bias may manifest itself in this way.

Bias may indicate that a different construct is being measured for different groups; whereas a test item may measure only achievement in algebra for one group, it might measure algebra and language comprehension for another group. This could lead to members of different groups receiving scores on the same test that are not truly

comparable. Bias may also indicate that an item may measure the same construct in different groups but with different levels of precision. In this case, an item might be quite good at assessing individuals' understanding of a concept in one group while it is an imprecise measure for members of another group. A result of this type of bias could be that test scores for members of different groups are not truly comparable because the same test more accurately measures a specific construct in one group than in another. Although bias is a technical term, it often leads to the socio-political construct of fairness (or lack thereof). In the context of the AP program, bias in testing can have important consequences in terms of what students are accepted into college and their college placement. If these high-stakes tests are biased, certain groups of students (e.g., females or African American students) will have important decisions made about their future on the basis of flawed information. Thus, there is a general consensus in the field of educational testing and research that bias is an important factor to be considered in both test development and evaluation (American Educational Research Association, 1999). The most commonly studied forms of bias are ethnicity- and sex-focused.

Ethnic bias in testing. There are a number of troubling examples of bias based on ethnicity within the context of traditional standardized testing programs. For example, (i) fewer minority students participate in advanced programs such as the AP program (e.g., Klopfenstein, 2004), (ii) minority students consistently perform worse on tests such as the SAT and AP exam (e.g., College Board, 2007), and (iii) minority students' scores on these exams are less robust predictors of success in college than the scores of White students (e.g., Noble, 2004).

Research has found that African American and Latino students enroll in AP courses at approximately half the rate of White students. Furthermore, this pattern holds even when controlling for the difference in the number of AP classes offered at predominantly White schools and predominantly minority schools (Klopfenstein, 2004). 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). Furthermore, the demographic breakdown of participants varies somewhat by subject area. For example, 66% of students who took the AP Physics B exam in 2006 were White whereas 6% were Latino and 3% were African American. By contrast, of students who took the English Language AP test, 12% were Latino and 7% were African American, and in Human Geography, 13% were Latino and 7% were African American (College Board, 2007). Since taking an AP course is a strong predictor of whether a student will take an upper level class or major in that subject (Dodd et al., 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 (see 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. This means that an African American student who scores at the 84th percentile when compared to other African American students would receive approximately the same score as a White student scoring at the 50th percentile among White students. 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 to 2.96 for White students (College Board, 2007). This difference is not only large but consequential: since 3 is a passing score, the average White student will pass an AP exam while the average African American student will fail.

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 to 2.95 for White students. But a difference of 1.13 separates the average scores of White students and African American on the AP Physics C exam, 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.

Next, the relationship between standardized test scores and freshman year GPA is somewhat weaker for African American and Latino students than for White students. Interestingly, a number of studies have shown that the SAT and ACT overpredict first-year GPA in underrepresented minorities (Noble, 2004; Sawyer, 1985; Young, 2004). That is, the freshman year GPA of these students is lower than would be expected based on their standardized test scores. This recurrent finding is counter to intuition: one would expect to see the overall lower standardized test scores of these students reflect their under-measured potential. Few studies have focused on explaining this finding and even fewer have been able to explain it adequately (see Zwick, 2007). Despite the lack of an empirically reasonable explanation for the direction of the effect, this is an important illustration of bias manifesting itself in differential predictive power.

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), and social-psychological differences (Steele, 1997). 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 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 and his colleagues (2006) designed augmented versions of the AP Psychology and AP Statistics examinations that included practical and creative subscales. The results of the study demonstrated that the actual AP scores were moderately positively correlated with scores on the augmented AP exams ($r = .61$ for Psychology, $r = .49$ for Statistics). Scores on each of the subscales were moderately positively correlated with scores on the actual AP examination (with r 's ranging from .33 to .54 for AP Psychology and from .36 to .45 for AP Statistics), with analytical and memory items exhibiting the highest correlations. As expected, the correlations suggested that the actual AP exams were more heavily weighted toward the assessment of analytical and memory skills. A key finding was that the effect-size difference between African American students and White students was virtually non-existent 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. Thus, it appears that not only do individual differences in profiles of strengths and weaknesses exist across cognitive skills (at least within the domains of psychology and statistics), but there is also some evidence to suggest the presence of systematic group differences as well. 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.

Sex bias in testing. Similar issues exist in terms of bias based on sex, starting with the persistent differences between the number of males and females enrolling in advanced high school courses in math, science, and computer programming (Stumpf & Stanley, 1996). The relatively low numbers of female students enrolling in these courses results in many more males taking AP exams in certain domains. These disparities have decreased substantially over the last ten years, but females are still very much underrepresented in certain subjects: they represent only 45% of AP Physics B test-takers, 22% of AP Physics C (Electricity and Magnetism), and 16% of AP Computer Science A and AB test-takers (College Board, 2007). This is not to say that male examinees outnumber females on all AP tests. For example, the majority of students who took AP exams in French Literature, Psychology, Spanish Literature, and Studio Art were female (College Board, 2007). Differences in participation based on sex are important for a number of reasons. Among them, the number of females

who take a particular AP exam is significantly positively related to the scores of females on that exam relative to males (Stanley, Benbow, Brody, Dauber, & Lupkowski, 1992; also see Stumpf & Stanley, 1998) and taking an AP exam in a certain discipline strongly increases the chances of that person pursuing a major and later a career in that field (Dodd et al., 2002; Morgan & Maneckshana, 2000). Thus, the educational and professional opportunities of females may be strongly affected by their participation (or lack thereof) in such advanced programs.

Past research has found significant differences in the performance of males and females on tests in different subject areas and on different types of items. Stanley and colleagues (1992) measured sex-related differences on 26 AP exams and 14 SAT II subject tests over a 3-year period and found that males scored significantly higher than females on AP Computer Science, Physics B, Physics C (Mechanics), and Chemistry exams, and SAT II exams in European History and Physics. On the other hand, females had slightly higher scores on AP exams in Latin, German, French Literature, English Literature, and Spanish Language, and SAT II exams in English Composition and English Literature. The effect-sizes of the differences between males' and females' scores tended to be smaller on the AP exams than on the SAT IIs; it has been suggested that the difference in effect-size is due to the fact that half of each AP exam is open-response and that females perform better on tests in which language is an important component (Stumpf & Stanley, 1996). Overall, females perform worse than comparable males on verbal SAT items about scientific topics or stereotypically male activities (e.g., sports or the military; Bridgeman & Schmitt, 1997). Females also perform worse on geometry problems but better on algebra

problems, and worse on applied word problems but better on “pure mathematics” questions than males (O’Neill & McPeck, 1993). In response to persistent findings related to sex bias, the Educational Testing Service began screening the SAT for items with possibly discriminatory content in 1989; however there has been effectively no change in the score gap between males and females since then (Burton & Burton, 1993).

Finally, the SAT, ACT, and most of the SAT IIs have been found to underpredict females’ college GPA by an average of 0.06 on a 4.00 grading scale (Willingham & Cole, 1997; Leonard & Jiang, 1999; Young, 2004). It has further been found that this difference is substantially reduced when writing scores from these assessments play a significant role in the predictive equation (Ramist et al., 1994; Leonard & Jiang, 1999). This suggests that tests involving a substantial writing component may be less prone to sex-based differences in predictive validity. As the latest version of the SAT includes a large writing component (College Board, 2008b), the underprediction of females’ college grades may be corrected in the future. There may be more cause for concern about the predictive validity of females’ scores on SAT IIs and certain AP exams that do not involve a significant writing component (e.g., those with highly quantitative content).

Biological, cultural, social psychological, and individual differences, as well as characteristics of the tests themselves, have all been suggested as explanations for the differences in the performance of males and females on standardized tests (see Wilder & Powell, 1989). The present research hypothesizes that these differences are related to the narrow range of cognitive processes that have been assessed by traditional tests

and that by testing for a broader range of processes, the achievement gap will be reduced. Although past research has found that the achievement gap between White and minority students is reduced by measuring a broader set of cognitive skills (e.g., Stemler et al., 2006), it remains to be seen whether such an augmentation will affect the achievement gap between males and females. Based on past research findings, including that the effect-size of the differences between the scores of males and those of females are smaller when tests have a large writing component (Stumpf & Stanley, 1996) and females do not perform as well as males on practical math problems but do better on less applied ones (O'Neill & McPeck, 1993), it seems reasonable to predict that the type of cognitive processing skills assessed has an important influence on measured ability.

Present research

The purpose of the current study was to examine the impact on student performance of creating a theory-driven assessment in the domain of physics that was based on a construct-validated theory of cognitive processing skills (Sternberg, 1997, 1999). This augmentation, which included the development of items to measure creative and practical skills rather than focusing exclusively on memory and analytical skills, represents a shift from measuring only content knowledge to measuring content knowledge and cognitive processes. This study aims to generalize the past successes using this procedure to other content domains. Physics was selected as the content area first because it is a “hard science” and thus this study has the potential to demonstrate the effectiveness of this procedure in a very different subject area than has previously been explored. Furthermore, whereas it is relatively

straightforward to imagine how test items in psychology or statistics can be creative or practical, physics may not intuitively lend itself to such an approach. Finally, physics is an important domain in which to address issues of achievement differences because females and minority students are not well represented in AP Physics classrooms and perform particularly poorly on the AP Physics exam compared to males and White students (College Board, 2007). The following research questions were of particular interest:

1. 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?
2. Does assessing a broad range of cognitive skills within the context of quantifying domain knowledge reduce differences based on ethnicity and sex in achievement as compared with assessments of memory and analytical skills only?
3. 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?
4. Is it possible to identify benchmarks that provide a better understanding of what students at different performance levels know and can do? If so, are these benchmarks the same for identifiable subgroups or do different subgroups know different things?

Method

Pilot study

A pilot version of the augmented AP Physics exam was created that consisted of 53 multiple-choice and 20 open-response items. The test was administered between March and May of 2006 to a total of 138 students.² The primary purpose of the pilot test was to identify items that performed poorly from a psychometric standpoint so that these items could be discarded or revised prior to the implementation of the main study. The results of the pilot test revealed that the multiple-choice section of the exam was too long given the time allotment since the total number of participants responding to each item declined sharply around item 45 and continued to decline until the end of the exam. Approximately one-third of test-takers did not complete the last items five items. This suggested that the optimal length for the multiple-choice section of the exam in the main study would be 45 items. Criteria for exclusion included negative item-total correlations and high difficulty values (as the test revealed too many difficult items relative to the ability level of the participants). In addition, the pilot analyses revealed limited variation in the scoring rubrics associated with the open-response items. Nearly all students received full credit on these items and, thus, they yielded very little information for the test developer. Finally, the results revealed a compelling need to develop more practical and creative items. According to three independent raters, only two items primarily tapped practical skills and only five items primarily tapped creative skills. As a result of these pilot analyses, a content area expert was retained to develop substitute items and the newly

² The data for the pilot study and the main study were collected as part of a research project funded by National Science Foundation Award #0710915; Principal Investigators: Robert Sternberg, Steven Stemler, and Elena Grigorenko.

developed items were rated by two independent raters with regard to the cognitive skills they assessed. The interrater reliability estimates exceeded .80. After the items had been revised and re-rated, the main study was conducted. The final version of the test that was used in the main study can be found in Appendix A.

Sample design and selection

A target sample of 20 AP physics teachers who would be willing to administer the augmented test to their students as a practice exam between March and May 2007 was sought.

In order to select the teachers who would be invited to participate, a list of all 2,383 teachers scheduled to administer the AP Physics B exam in the spring of 2007 was acquired from the College Board. Out of these, 146 teachers were teaching outside of the United States and its territories and were excluded from the pool of potential participants. Teachers in United States territories and commonwealth associates (e.g., Guam, Puerto Rico, and the U.S. Virgin Islands) were also excluded, leaving only the 2,231 teachers administering the exam in the continental United States and Hawaii. These 2,231 teachers were grouped into 9 regions, following the United States Department of Education's regional breakdown into 10 Regional Education Laboratories. The *Northwest* and *Pacific* regions were combined, since the latter only included the state of Hawaii once territories had been removed.

Demographic information for each school was added to the existing AP Physics teacher database. The demographic information included the absolute numbers of African American, Latino, and White students enrolled in each school, based on information obtained from the National Center for Education Statistics

(<http://nces.ed.gov/ccd/schoolsearch/>). Using this information, a concerted effort was made to maximize the ethnic diversity of the sample. The schools in each region were ranked in decreasing order of the number of African American students and Latino students. Two lists were generated for each region. The first was a list of teachers from schools with the most African American students and schools with the most Latino students. The second list consisted of teachers from schools with the highest combined numbers of Latino and African American students. Teachers who appeared on both of the lists were included in the final selection of teachers to be recruited. For regions where there was insufficient overlap between lists, additional teachers from the schools with the highest numbers of both African American and Latino students were included in the final selection of teachers to be recruited.

In the first wave of recruitment, invitation letters were sent to 35% of teachers in each region, for a total of 62 teachers. Within each region, half the potential participants were randomly selected from teachers who had been teaching for 0-5 years and half among teachers who had been teaching for 6 years or more. For example, in the *Northeast* region, there were a total of 358 teachers. Out of these teachers, 10 were randomly selected to receive invitations, 5 of whom had been teaching for 5 years or less, and 5 of whom had been teaching for 6 years or more.

After sending out invitations to 62 randomly selected teachers from ethnically diverse schools, positive responses were received from 9 of them, teaching a total of 298 students. In order to complete the sample and reach the desired number of 20 teachers and 300 or more students, another 62 teachers were randomly selected using the same procedure that was used for the first wave of invitations. The 62 teachers

randomly selected for the first wave of invitations were removed from the database before the second random sampling. After sending out the second wave of invitations, positive replies were received from 7 teachers, teaching a total of 141 students. However, not all of the teachers were able to comply with the time and date demands of the project, so a third and final wave of invitations was sent to another 62 randomly selected teachers after having removed the teachers who responded to waves one and two from the database.

Final sample

A total of 10 teachers from 10 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 female and 26 students who failed to indicate their sex. In terms of ethnicity, 87 students were White, 69 were Asian or Pacific Islanders, 40 were Latino, 32 were African American, and 5 reported multiple ethnicities, with the remaining 48 students not responding.

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 hours. 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 using a Scantron machine. Open-response sections were scored by a group of independent raters who were content experts. These raters were provided with guidelines for scoring the open-response items; a copy of these guidelines and an answer key for the multiple-choice section can be found in Appendix B.

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 to reliably discriminate between people of different ability levels. Next, the factor structure, 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. Information about how different groups performed on the test is presented next. This includes analyses of differential test functioning based on ethnicity and sex (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). Finally, an effort is made to better understand the meaning behind group differences by looking at the items that distinguished between groups of different performance levels overall as well as by ethnicity, sex, and profile type (i.e., benchmarks).

As a precursor to these analyses, it was necessary to accurately and reliably categorize items based on whether they represented memory, analytic, creative, or practical processes. First, three content experts were trained in identifying process types. The number of items that were categorized as primarily falling into each of the four processes by these raters appears below.

Table 1

Number of items categorized as primarily (>51% of skill tapped) falling into each process area

	Primarily memory	Primarily analytic	Primarily creative	Primarily practical	No primary skill	Total items with a primary skill
Rater 1	27	94	0	57	86	178
Rater 2	11	130	20	49	54	210
Rater 3	43	87	57	66	11	253

Although they showed reasonably consistent agreement (the pattern of agreement ratings ranged from the .50s to the high .70s), the ratings of these content experts

yielded almost no variability. Raters 1 and 2 believed that nearly two-thirds of the items they rated as primarily tapping one type of thinking required memory or analytical skills, and Rater 3 believed that about half of the items tapped these two domains. All raters agreed that analytical thinking was the dominant category and identified the majority of items as analytic.

Two process experts who had little to moderate knowledge of physics rated the items as well; because their ratings were highly reliable and yielded more variability than those of content experts, their classifications were used to construct the process subscales that are employed in future analyses. Although the ratings of these process experts exhibited very high consensus reliability (Cohen's kappa = .71), to be conservative, only those items where both raters agreed on the primary 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 be included in the creative subscale. Out of 69 items, 54 met this criterion; 37 were multiple-choice items and 17 were open-response items. A list of items associated with each cognitive process area appears below.

Table 2

Summary of final process subscales

Memory process subscale	Analytic process subscale	Creative process subscale	Practical process subscale
14	2	4	1
25	5	8	3
27	9	13	10
37	11	17	15
.	12	30	16
.	18	34	20
.	19	38	21
.	23	44	24
.	26	Problem 4, Question 3	33
.	28	.	39
.	29	.	40
.	35	.	Problem 1, Question 3
.	45	.	Problem 2, Question 4
.	Problem 1, Question 1.1	.	Problem 3, Question 3.2
.	Problem 1, Question 1.2	.	Problem 3, Question 4
.	Problem 2, Question 2	.	Problem 4, Question 4
.	Problem 3, Question 1.1	.	Problem 5, Question 3
.	Problem 3, Question 2.1	.	.
.	Problem 3, Question 2.2	.	.
.	Problem 3, Question 3.1	.	.
.	Problem 4, Question 1	.	.
.	Problem 5, Question 1	.	.
.	Problem 5, Question 2	.	.

Item statistics

Out 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. On the other hand, 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. The average score on each open-response item ranged from 0.19 to 2.89, with possible point totals ranging from 2 to 6.

Two items (40 and 26) had negative item-discrimination values (also called item-total correlations), indicating that performance on these items was negatively related to overall performance on the test; however, as the values were close to zero, they were not a major concern. 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. Just over a quarter of the items (29%) had item discrimination values of .30 or above when looking at the relationship between performance on each item and performance on the overall test; however, when looking at each item and performance on the subsection (multiple-choice or open-response), there was a larger proportion of items with discrimination values of .30 or above (38%). Overall, it appears as though getting a single item correct was reasonably related to getting a high score on the test as a whole. Item difficulty and discrimination statistics are listed by item in Appendix C.

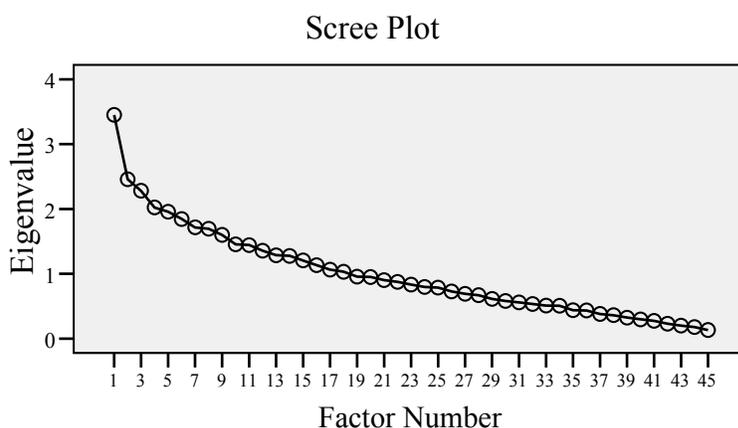
Test statistics

Factor analysis. The multiple-choice and open-response sections of the test were analyzed separately in factor analyses. Analyzing the sections together would likely have resulted in a factor structure that reflected the differences in response format more than anything else. As each item already contained process and content

information, the decision was made not to introduce an additional dimension on which items might vary (i.e., multiple-choice versus open-response format).

The factor analytic results of the multiple-choice section suggested that each item contributed unique variance into the equation. When eigenvalues over 1 were extracted (using principal axis extraction with a promax rotation³), 18 distinct factors emerged that explained a combined 47.2% of the variance in the dataset. Inspection of the scree plot revealed that the marginal percent of variance explained by each additional factor dropped sharply after four factors (see Figure 1). When four factors were extracted, 15.6% of the total variance in the dataset was explained with the first factor explaining only 5.9% on its own. These factors did not meaningfully map on to content or process categories.

Figure 1. Scree plot for the multiple-choice section factor structure.



For the open-response section, a principal axis analysis with promax rotation set to extract factors with eigenvalues greater than 1 yielded an 8-factor solution that

³ Principal axis extraction was selected because it only considers common variance rather than unique variance (i.e., error) in a set of variables. This method of extraction is generally preferred when the goal is insight into data structure rather than data reduction (Widaman, 1993). A promax rotation was chosen because it allows factors to be correlated with one another and this research does not intend to suggest that the characteristics of the items (i.e., content or process) are completely unrelated to each other.

explained 48.8% of the variance in the data, 18.3% of which was explained by the first factor. Except for the first factor, the factors seemed to correspond more or less to each of the main problems on this portion of the exam (or in some cases, subsections of the main problems).

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 reasonably high for 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).

Rasch measurement statistics. The Rasch model (Rasch, 1960/1980) offers a way to understand item responses as a function of person and item parameters. According to this model, the probability of any given correct response can be represented as a logistic function of the difference between person ability and item difficulty as follows

$$\log[p_{ni}/(1-p_{ni})] = B_n - D_i \quad (1)$$

where B_n is the ability of a person n and D_i is the difficulty of the item i . Thus, the probability of a correct response will be higher as the person's ability increases relative to item difficulty. The following results address the degree to which this assessment conformed to the expectations of the Rasch model.

The Rasch model places person ability and item difficulty on the same (logit) scale such that they can be directly compared. This is accomplished by taking the success-to-failure ratio for each person (i.e., the percentage of items they answered

correctly) and item (i.e., the percentage of people who answered that item correctly) and converting these scores to their natural log odds (Bond & Fox, 2001). The results of these computations can be represented on an item-person map, wherein person ability estimates are ordered on one side of a logit scale and item difficulty estimates are ordered on the other side, allowing for a visual representation of how difficult items on a test are relative to the ability of the test-takers. Because both item difficulties and person ability estimates have been converted to a logit scale, items and people in equivalent locations are comparable. That is, an item with a difficulty estimate of 0.50 logits should be at the exact ability level of a person with an ability estimate of 0.50 logits. Figure 2 shows a map of item difficulties relative to person abilities. Here, the abilities of individual test-takers are represented on the left side of the figure (from the highest ability to lowest ability); each # symbol represents four people. Item difficulties are represented on the right side of the figure (from most difficult to least difficult). For items that were associated with a process subscale, the item number is preceded by an italicized M (for memory), A (for analytic), C (for creative), or P (for practical), followed by an underscore.

independent of item difficulty. That is, it was not the case that all items that corresponded to a particular process (i.e., analytic items) were harder or easier than other items; rather, items on the same subscales were spread out in terms of difficulty level.

The Rasch model offers two types of fit statistics that indicate how well each item on a test fits the assumption of a unidimensional construct: outfit and infit mean square statistics. Outfit mean square statistics are based on the mean sum of squared standardized residuals for each person and item. Infit mean square statistics are based on the squared standardized residuals for each person and item, which are weighted with respect to the variance of the observation (with larger weights assigned to well-targeted observations and smaller weights to extreme ones) and then summed (Bond & Fox, 2001; Wright & Masters, 1982; Wright & Stone, 1979). Thus, infit statistics give more weight to items that are well-matched to the test-takers' abilities and people who are well-matched to the items on the assessment. Because person ability and item difficulty were not well-matched on the present assessment, infit mean square statistics are reported. 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 to 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 expects some low-ability people will answer difficult items correctly and some high-

ability people will answer 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.

Next, the Rasch model offers person and item reliability estimates that can be interpreted in the same way as the 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 (Wright & Masters, 1982). 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 in 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 a large group of below-average students.

Group statistics

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

Ethnic differences in ability. As reported previously, of the 281 students who took the exam, 87 were White, 69 were Asian or Pacific Islanders, 40 were Latino, 32 were African American, 5 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 ($n = 37$) and individual person ability estimates were calculated separately for each ethnic group based on the Rasch model. As discussed in the previous section, an ability estimate of 0.00 logits indicates that the level of person 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), and 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 answered). The range of ability estimates as well as the means and standard deviations by ethnic group appear below.

Table 3

Descriptive statistics of abilities by ethnic group

		Range	Mean	SD
White students	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	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	-0.73	0.41
Latino students	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 students	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

On the overall test, Asian students had the most variability in their ability estimates.

Asian and White students showed more varied ability estimates on the memory subscale than other ethnic groups, and Asian and Latino students showed the more varied ability estimates on the analytic and practical subscales. White students exhibited the least varied ability estimates on the creative subscale compared to other groups. Across all subscales, Asian students demonstrated a high degree of variability

in their ability estimates, but African American students' ability estimates were far less varied, especially on the practical subscale. African American students also had the lowest mean ability estimates, followed by Latino, Asian, and White students. Each ethnic group had the highest mean ability estimates on the memory subscale, followed by the practical subscale, the creative subscale, and the analytic subscale. The observed differences in mean ability estimates of White students compared to non-White students were examined looking at significance levels and effect-sizes; the results of these analyses appear in Table 4 below. Although there are many combinations of ethnic group differences that could be examined, this research will only focus on the differences in achievement of White students compared to non-White students. This is in part because White students comprise the majority of test-takers on the actual AP Physics exam and in this sample, so their performance (whether correctly or not) is often considered the baseline. Another reason is that the differences in achievement reported by past research have typically compared the achievement of minority students to the achievement of White students. Thus, in order to compare the effectiveness of this assessment in reducing standard differences in achievement based on ethnicity, it is necessary to compare the performance of minority groups to that of White students.

Table 4

Differences in ability based on ethnic group

		<i>t</i> -statistic	Significance	Effect-size (Cohen's <i>d</i>)
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 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

** Difference is significant at the 0.01 level (2-tailed).

* Difference is significant at the 0.05 level (2-tailed).

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 analytic 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 students did not perform significantly differently on the test as a whole. The only subscale on which White and Asian students showed a reliable difference in measured ability was on the

practical subscale, where Asian students performed worse than White students by a relatively large amount.

Sex differences in ability. The same methods used to examine ethnic differences were used to identify differences between the measured ability of males and that of females. The data were split into two groups based on the sex reported by the test-taker; of the 281 students who took the exam, 151 were males, 104 were females, and the remaining 26 failed to report sex, and were therefore excluded from the following analyses. Item difficulties for each group were anchored based on the item difficulties generated by the set of participants who responded to every item and person ability estimates were calculated separately for males and females. Descriptive statistics of male and female abilities appear in Table 5 below.

Table 5

Descriptive statistics of abilities by sex

		Range	Mean	SD
Males	Whole test	2.64	-0.36	0.47
	Memory subscale	5.71	0.29	1.47
	Analytic subscale	7.43	-0.76	0.98
	Creative subscale	5.71	-0.37	1.08
	Practical subscale	4.15	-0.27	0.54
Females	Whole test	2.06	-0.56	0.42
	Memory subscale	5.71	0.01	1.48
	Analytic subscale	4.68	-0.79	0.71
	Creative subscale	4.20	-0.70	0.98
	Practical subscale	4.50	-0.60	0.62

For the most part, males showed more variability in their ability estimates than females, especially on the analytic subscale. Females showed slightly more variation in ability estimates on the practical subscale. Both groups showed the highest ability estimates on the memory subscale, followed by the practical, creative, and analytic

subscales. In addition to performing better on the test as a whole, males received higher ability estimates on each subscale, although the difference between male and female ability estimates was the smallest on the analytic subscale. These differences were examined for significance and effect-size, the results of which appear in Table 6 below.

Table 6

Differences in ability based on sex

		<i>t</i> -statistic	Significance	Effect-size (Cohen's <i>d</i>)
Females v. Males	Whole test	3.55	.00**	-0.45
	Memory subscale	1.51	.13	N/A
	Analytic subscale	0.25	.80	N/A
	Creative subscale	2.44	.02*	-0.31
	Practical subscale	4.49	.00**	-0.56

Note. Cohen's *d* statistics are computed using male students as the reference group

** Difference is significant at the 0.01 level (2-tailed).

* Difference is significant at the 0.05 level (2-tailed).

Males performed significantly better than females on the test as a whole. This difference was moderate as were the differences on the creative and practical subscales. There were no significant differences in the measured abilities of males compared to females on the analytic or memory subscales.

Profile group differences in ability. Whereas the most common forms of factor analysis load factors across variables, reducing the variables into meaningful categories, Q-type factor analysis loads factors across individuals, reducing participants into identifiable groups. Using principal components extraction with a promax rotation, three factors with eigenvalues greater than 1.0 were obtained, which accounted for 100% of the variance in students' performance patterns. The three factors corresponded to distinct profiles of achievement represented in the dataset.

Table 7 presents an abridged structure matrix of factor loadings for each participant on the exam. It is important to note that while some participants had positive loadings on a factor, others had negative loadings on the same factor. The participants with negative loadings on a factor demonstrated the opposite pattern of achievement from those participants with positive loadings on the factor. Therefore, although three factors were extracted, they yielded six distinct profiles of achievement.

Table 7

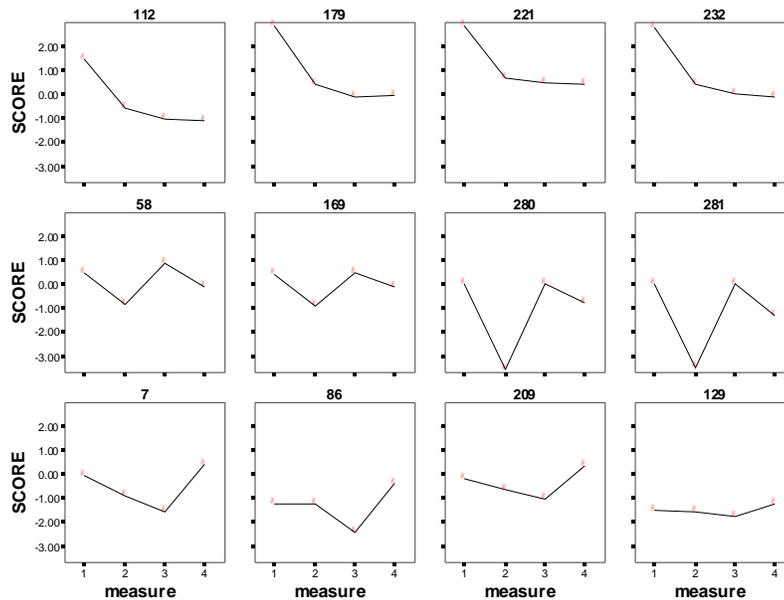
Abridged output of factor loadings for each participant

	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

To demonstrate the meaning of each of these factors, Figure 3 depicts the profiles of achievement for 12 participants. The participants in the first row had high positive loadings on the first factor; the participants in the second row had high positive

loadings on the second factor; and the participants in the third row had high positive loadings on the third factor.

Figure 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.

Participants with high loadings on the first factor tended to exhibit high performance on the memory subscale with relatively lower scores on the analytic, creative, and practical subscales. Their performance on the analytic subscale was slightly better than their performance on the creative or practical subscales. Participants with negative loadings on this factor exhibited the opposite pattern of performance; they performed poorly on the memory subscale and better on the analytic, creative, and practical subscales.

Participants with high loadings on the second factor demonstrated high performance on memory and creative subscales, lower performance on the practical subscale, and very low performance on the analytic subscale. Participants with

negative loadings on this factor showed a reverse pattern of performance; these participants performed poorly on the memory and creative subscales, better on the practical subscale, and best on the analytic subscale.

Lastly, participants with high positive loadings on the third factor had high scores on the practical subscale, low scores on the creative subscale, and moderate scores on the memory and analytic subscales. Participants with negative loadings on this factor had low scores on the practical subscale and high scores on the creative subscale.

At this point, it is essential to note that these patterns do not correspond to absolute achievement differences. Rather, a student who performs well on the test as a whole and one who performs poorly could exhibit the same achievement profile. For example, a high-performing student's weakest area may be practical processing (even though that student is stronger than most other students in this area, it is her particular area of weakness) and strongest area may be memory processing. Similarly, a low-performing student's weakest area be practical processing and best area memory processing (even though he performs worse than most other students in this area, it is his best area of the four). Thus, these results reflect patterns of achievement across cognitive process skills and should not be confused with absolute achievement. To further illustrate this point, descriptive statistics of each profile group's measured ability on each subscale and on the test as a whole appear below in Table 8.

Table 8

Descriptive statistics of ability by profile group

		Range	Mean	SD
Profile 1 (positive loading)	Whole test	2.32	-0.35	0.46
<i>High memory (HM)</i>	Memory subscale	4.32	1.29	1.12
	Analytic subscale	3.26	-0.48	0.65
	Creative subscale	4.86	-0.50	0.91
	Practical subscale	4.55	-0.48	0.67
Profile 1 (negative loading)	Whole test	2.36	-0.50	0.43
<i>High analytic, high practical, high creative (HAHPHC)</i>	Memory subscale	4.35	-1.43	1.19
	Analytic subscale	2.79	-0.68	0.60
	Creative subscale	2.60	-0.58	0.63
	Practical subscale	3.55	-0.35	0.57
Profile 2 (positive loading)	Whole test	3.61	-0.62	0.57
<i>High memory, high creative (HMHC)</i>	Memory subscale	4.32	0.03	0.80
	Analytic subscale	6.22	-1.53	1.18
	Creative subscale	3.72	0.15	0.67
	Practical subscale	3.04	-0.44	0.55
Profile 2 (negative loading)	Whole test	1.76	-0.50	0.53
<i>High analytic, high practical (HAHP)</i>	Memory subscale	2.44	-1.11	1.06
	Analytic subscale	2.85	-0.19	0.91
	Creative subscale	2.42	-1.97	0.94
	Practical subscale	2.44	-0.73	0.81
Profile 3 (positive loading)	Whole test	2.34	-0.53	0.40
<i>High memory, high practical (HMHP)</i>	Memory subscale	4.35	-0.41	1.00
	Analytic subscale	3.68	-0.96	0.70
	Creative subscale	3.40	-1.60	0.97
	Practical subscale	2.38	-0.19	0.44
Profile 3 (negative loading)	Whole test	2.08	-0.43	0.49
<i>High creative (HC)</i>	Memory subscale	2.48	-0.45	0.70
	Analytic subscale	3.22	-0.41	0.78
	Creative subscale	4.39	0.46	0.97
	Practical subscale	2.22	-0.68	0.52

As expected, mean ability estimates for each profile group were the highest on the group's strongest subscales and lowest on the group's weakest subscales. Although this finding should be intuitive from the profiles of exemplary achievement, it is important to note that not only did particular individuals within a profile group

exhibit the same strengths and weaknesses as each other, but this held over the aggregate as well. Participants who were characterized by comparatively strong memory skills tended to perform better on the test overall, as evidenced by the high average ability estimate associated with profile 1 (positive loading). Participants associated with profile 3 (negative loading) also had high ability estimates based on the assessment as a whole; these participants exhibited relative strength in creative processes and relative weakness in practical processes. The group with the lowest ability estimates (profile 2 – positive loading) exhibited strong memory and creative skills, and poor analytical and practical skills.

Table 9 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 (profile 1 – positive loading), but many participants exhibited profiles that were not characterized by strong memory or analytic skills.

Table 9

Summary of participants associated with each profile

	Frequency	Percent
Profile 1 - <i>HM</i>	107	38.10%
Profile 1 - <i>HAHPHC</i>	41	14.60%
Profile 2 - <i>HMHC</i>	61	21.70%
Profile 2 - <i>HAHP</i>	9	3.20%
Profile 3 - <i>HMHP</i>	40	14.20%
Profile 3 - <i>HC</i>	23	8.20%

A chi-square test of association revealed that the actual number of participants that were associated with each profile divided by sex and ethnicity were 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 10 and 11 summarize the number of participants associated with a particular profile of achievement broken down by ethnicity and sex, respectively.

Table 10

Summary of participants associated with each profile by ethnicity

		Ethnicity					
		African American	White	Latino	Asian	Total	
Profile	Profile 1 - <i>HM</i>	Count	10	41	12	25	88
		% of total	4.4%	18.0%	5.3%	11.0%	38.6%
	Profile 1 - <i>HAHPHC</i>	Count	8	13	6	11	38
		% of total	3.5%	5.7%	2.6%	4.8%	16.7%
	Profile 2 - <i>HMHC</i>	Count	7	16	9	12	44
		% of total	3.1%	7.0%	3.9%	5.3%	19.3%
	Profile 2 - <i>HAHP</i>	Count	1	2	1	2	6
		% of total	0.4%	0.9%	0.4%	0.9%	2.6%
	Profile 3 - <i>HMHP</i>	Count	5	10	9	13	37
		% of total	2.2%	4.4%	3.9%	5.7%	16.2%
	Profile 3 - <i>HC</i>	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%

Table 11

Summary of participants associated with each profile by sex

Profile		Sex		Total
		Male	Female	
Profile 1 - <i>HM</i>	Count	61	42	103
	% of total	23.8%	16.4%	40.2%
Profile 1 - <i>HAHPHC</i>	Count	27	14	41
	% of total	10.5%	5.5%	16.0%
Profile 2 - <i>HMHC</i>	Count	28	20	48
	% of total	10.9%	7.8%	18.8%
Profile 2 - <i>HAHP</i>	Count	4	3	7
	% of total	1.6%	1.2%	2.7%
Profile 3 - <i>HMHP</i>	Count	21	17	38
	% of total	8.2%	6.6%	14.8%
Profile 3 - <i>HC</i>	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%

Benchmarking

The process of benchmarking, or identifying a set of items that reliably separate groups of different achievement levels, may yield important information regarding the progression of physics understanding, allowing for insight into what individuals and groups at various performance levels know and do not know about physics. The 90th, 75th, 50th, and 25th percentiles of total scores were computed and groups were created based on the percentile range into which participants fell. For example, the 90th percentile group consisted of all participants who scored at the 90th percentile mark or above on the test as a whole. For each benchmark group, item difficulties were calculated and potential benchmark items were selected based on whether strictly more than 65% of the group in question answered the item correctly and

strictly less than 50% of the next lower group answered the item correctly.⁴ Items were weighted based on how many points they were worth such that the difficulty of an open-response item worth 3 points would be divided by 3 prior to assessing if it met the benchmark criteria. When the procedure yielded multiple candidates, the items with the largest discrepancy between group difficulties were selected to serve as benchmarks in comparisons across groups.

One multiple-choice and one open-response item were selected as benchmarks to distinguish the 90th, 75th, and 50th percentile groups of test-takers. No items met the criteria to serve as benchmarks identifying the 25th percentile, indicating that the top 50-75% of test-takers did not differ reliably from the bottom 25% of test-takers. All of the items that emerged as suitable benchmarks are listed below in Table 12 with those selected to serve as benchmark items in group comparisons appearing in bold.

Table 12

Benchmark items

Benchmark items		
90 th percentile	14 ¹	Problem 3, Question 4 ⁴
75 th percentile	23, 32 , 35	Problem 3, Question 3.2 ⁴
50 th percentile	11 ² , 22, 31	Problem 1, Question 3 ⁴
25 th percentile	.	.

¹ Memory subscale² Analytic subscale³ Creative subscale⁴ Practical subscale

Out of five benchmark items from the overall group that were associated with a process subscale, three of them (all open-response items) were from the practical

⁴ This procedure was based on that used by Kelly, Mullis, & Martin (2000) to identify international benchmarks on the Third International Mathematics and Science Study (TIMSS).

subscale. For information regarding the content of these items, please refer to Appendix A, where a copy of the entire test can be found.

An individual was said to have achieved the 90th percentile benchmark if she answered the two benchmark items for that level (Item 14 and Problem 3, Question 4) correctly.⁵ Note, that this does not necessarily mean that this student was in the 90th percentile in terms of her overall score, merely that she met the benchmark associated with the 90th percentile. Similarly, a student was said to have achieved the 75th and 50th percentile benchmarks if she answered the benchmark items associated with these levels correctly. Based on these criteria, a student who answered all six benchmark items correctly was labeled as meeting the 50th, 75th, and 90th percentile group benchmarks. Thus, each student was associated with three binary variables that indicated whether or not they met the 50th, 75th, and 90th percentile benchmarks. Chi-square analyses showed that there were significant differences in the ethnic and sex composition of the students who met each of these percentile benchmarks. The results of these analyses are summarized in Table 13 below.

⁵ For open-response items, a correct response was defined as receiving at least half credit on the item.

Table 13

Summary of ethnic and sex differences in achieving benchmarks

	Ethnicity			
	χ^2	<i>df</i>	<i>N</i>	<i>p</i>
90 th percentile	4.58	3	228	.21
75 th percentile	8.01	3	228	.05*
50 th percentile	11.82	3	228	.01**
	Sex			
	χ^2	<i>df</i>	<i>N</i>	<i>p</i>
90 th percentile	3.96	1	256	.05*
75 th percentile	8.61	1	256	.00**
50 th percentile	0.44	1	256	.51

** Difference is significant at the 0.01 level (2-tailed).

* Difference is significant at the 0.05 level (2-tailed).

Specifically, significant differences were found in the ethnic composition of the group of students who met the 75th and 50th percentile benchmarks, although not in the group who met the 90th percentile benchmark. White students were more likely to meet the 75th percentile benchmark than African American students; 26% of White students achieved this benchmark compared to 6% of African American students. White students were also more likely to meet the 50th percentile benchmarks than either African American or Latino students; 43% of White students reached this benchmark compared to 23% of Latino students and 13% of African American students. Differences in the sex breakdown of the groups were significant for the 90th and 75th percentile groups, although not the 50th percentile group. This was the result of more males meeting the 90th and 75th percentile benchmarks than females; 11% of males met the 90th percentile benchmark compared to 4% of females, and 24% of males met the 75th percentile benchmark compared to 10% of females.

Benchmarks based on ethnic groups. It is possible that different ethnic groups have different benchmarks. That is, the items that effectively distinguished the top

10% of students overall may not be the same items that distinguished the top 10% of African American students. This information has the potential to shed light on the group differences that were found on the test. Therefore, the benchmarking process described above was repeated for each ethnic group. The following benchmarks represent the items that effectively distinguished between the performance levels of White, African American, Latino, and Asian students. As before, although all potential benchmarks are listed, those that appear in bold represent the items that were ultimately selected to serve as benchmarks in group comparisons.

Table 14

Benchmark items based on ethnicity

Benchmark items based on White students		
90 th percentile	7	Problem 5, Question 4
75 th percentile	14 ¹ , 32	Problem 4, Question 1²
50 th percentile	21 ⁴ , 35 ²	Problem 3, Question 3.2⁴
25 th percentile	27 ¹ , 28 ² , 31	Problem 1, Question 1²
Benchmark items based on African American students		
90 th percentile	7, 10 ⁴ , 12 ² , 15 ⁴ , 16 ⁴ , 17 ³ , 25 ¹ , 27 ¹ , 40 ⁴	.
75 th percentile	11 ² , 33 ⁴	Problem 1, Question 1.2 ²
50 th percentile	.	.
25 th percentile	20 ⁴ , 21 ⁴	.
Benchmark items based on Latino students		
90 th percentile	2 ² , 8 ³ , 14 ¹ , 25 ¹ , 31, 32, 41	Problem 3, Question 4⁴
75 th percentile	13 ³ , 26 ² , 28 ² , 36	Problem 2, Question 4⁴
50 th percentile	7, 38 ³	Problem 1, Question 3⁴
25 th percentile	2 ² , 24 ⁴ , 31, 41	Problem 1, Question 1.1²
Benchmark items based on Asian students		
90 th percentile	7, 11 ² , 24 ⁴ , 28 ² , 33 ⁴ , 35 ² , 45 ²	Problem 3, Question 1.2; Problem 3, Question 4 ⁴ ; Problem 4, Question 3³
75 th percentile	17 ³ , 21 ⁴ , 23 ² , 32	Problem 5, Question 4
50 th percentile	22, 36	Problem 1, Question 3⁴
25 th percentile	33 ⁴	Problem 1, Question 1.2^{2*} ; Problem 2, Question 1; Problem 5, Question 4

¹ Memory subscale² Analytic subscale³ Creative subscale⁴ Practical subscale*Problem 5, Question 4 exhibited a higher discrepancy between groups but it was not selected because it is being used as the 75th Percentile benchmark.

Although no items could reliably distinguish between the 25th percentile group and those students who scored below the 25th percentile mark for the overall group, when the items were analyzed separately based on ethnicity, there were items that successfully distinguished between these groups. However, among African American students, there were no items that reliably separated the 50th percentile group from the

surrounding groups. Additionally, only one open-response item was capable of distinguishing between any of the percentile groups among African American students. As such, two multiple-choice items were selected as benchmarks for each level for this group (instead of one multiple-choice item and one open-response item) to allow comparisons to be made across percentile groups. Strangely, there were four cases in which the same item was a possible benchmark at multiple levels in a single ethnic category: Items 2, 31, and 41 marked both the 90th and the 25th percentile groups for Latino students, and Problem 5, Question 4 marked both the 75th and the 25th percentile groups for Asian students.

The benchmarks for White and Asian students were most similar to those for the overall group. Out of 12 potential benchmarks for White students, 3 of them (25%) were also benchmarks for the overall group. Similarly, Asian students shared 6 out of 22 (27%) potential benchmarks with the overall group. Latino students had benchmarks that were slightly less similar to the overall group: out of 20 potential benchmarks, Latino students and the overall group had 4 (20%) in common. Finally, African American students' benchmarks were markedly different from those of the group as a whole. Only 1 out of 14 (7%) potential benchmarks was shared.

Less than half of the potential benchmarks for White students were benchmarks for other groups. White students and Asian students shared five benchmark items: Item 7 marked the 90th percentile for both White and Asian students; Item 28 marked the 90th percentile for Asian students and the 25th percentile for White students; Item 32 marked the 75th percentile for both White and Asian students; Item 35 marked the 90th percentile for Asian students and the 50th percentile for White students; and

Problem 5, Question 4 marked the 90th percentile for White students and both the 75th and the 25th percentiles for Asian students. Two of these benchmarks served to separate comparable groups in both White students and Asian students (both multiple-choice items), but two separated the highest scoring Asian students from their lower scoring peers and much lower scoring White students from their peers (also, both multiple-choice items), and one separated the highest scoring White students and lower scoring Asian students (an open-response item).

Although five benchmark items were shared by White students and Latino students, few of the items separated comparable groups. Item 7 marked the 90th percentile for White students but the 50th percentile for Latino students; Item 14 marked the 75th percentile for White students but the 90th percentile for Latino students; Item 28 marked the 25th percentile for White students but the 75th percentile for Latino students; Item 31 marked the 25th percentile for White students but the 90th and the 25th percentile for Latino students; and Item 32 marked the 75th percentile for White students but the 25th percentile for Latino students. In most cases, these shared items were successful in distinguishing high performing Latino students and lower performing White students from their respective peers.

Finally, only two benchmark items were the same for White students and African American students: Item 7 marked the 90th percentile for both White students and African American students, and Item 21 marked the 50th percentile for White students but the 25th percentile from African American students. The smaller number of matching benchmarks certainly followed from the lack of successful open-response benchmarks and the impossibility of identifying items that distinguished between the

50th and 25th percentile groups among African American students. Overall, few benchmark items were shared across ethnic groups and fewer were shared by corresponding percentile groups.

Attention should also be paid to the cognitive processes represented by these benchmarks. Out of 48 benchmarks for all ethnic groups associated with process subscales, 17 (35%) were rated as practical items. For White students, practical items were useful in distinguishing the 50th percentile group in particular whereas analytic and memory items were useful benchmarks at every level. Among African American students, practical items were reliable benchmarks at the lowest level of performance; both practical and analytic items marked the 75th percentile group; and all four processes were represented by the benchmarks that separated the 90th percentile group. Practical items were also useful benchmarks at every level for Latino test-takers. For Latino students, the 50th and 75th percentile group benchmarks also contained creative items and the 90th percentile group benchmark included memory items. For Asian students, practical items served as benchmarks at every level whereas memory items were never present as benchmarks.

After identifying benchmark items for each group and examining the differences and similarities between benchmarks, the question of who met these benchmarks was addressed. As before, binary variables were created to represent whether or not each individual met each ethnicity-specific benchmark. For example, each student was coded as achieving the 90th percentile benchmark based on African American students if he answered Items 7 and 16 correctly (the benchmarks items based on African American students at this level) but as failing to achieve this benchmark if he

answered these items incorrectly. Similarly, all participants who answered Item 7 correctly and received at least half credit on Problem 5, Question 4 (the 90th percentile benchmarks based on White students) were coded as achieving the 90th percentile benchmark based on White students. Again, it is important to emphasize that if a student met the 90th percentile benchmark based on White students or based on African American students it does not follow that he scored in the 90th percentile of either of these groups, only that he answered the benchmark items at that level correctly.

The correlations between meeting each benchmark based on the overall group and each ethnic group were calculated using a phi correlation (a type of correlation used when both variables are dichotomous), controlling for total score. The goal behind this was to determine if a basic knowledge was shared by different groups at the same performance level that was independent of overall performance. For example, an African American student who met the 90th percentile benchmark based on his own ethnic group may be more likely to meet the overall 90th percentile benchmark simply because he answered a large number of items correctly. But controlling for his overall performance ensures that any observed relationship is the result of shared features of the benchmark items and shared characteristics of this student and other students who met the overall 90th percentile benchmark. Table 15 presents the relationship between meeting each overall benchmark and the corresponding ethnic group benchmark.

Table 15

Correlations between ethnic group benchmarks and overall benchmarks

		90 th percentile benchmark (overall)	75 th percentile benchmark (overall)	50 th percentile benchmark (overall)
Corresponding White benchmark	Partial correlation	.08	.07	.04
	Significance	.19	.25	.49
Corresponding African American benchmark	Partial correlation	-.09	.03	N/A
	Significance	.14	.60	N/A
Corresponding Latino benchmark	Partial correlation	.10	-.09	.19**
	Significance	.11	.15	.00
Corresponding Asian benchmark	Partial correlation	.03	.45**	.37**
	Significance	.63	.00	.00

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

As can be seen above, there were significant relationships between meeting the 75th percentile benchmark overall and based on Asian students, and meeting the 50th percentile benchmark overall and based on Asian and Latino students. There were no other significant relationships between ethnic group benchmarks.

Because White students represented the majority of test-takers (both in this sample and on most AP exams), it is also important to look at how White students' benchmarks differed from the benchmarks based on other ethnic groups. Phi correlations between meeting each benchmark based on White students and each other ethnic group (controlled for total score) are reported in Table 15 below.

Table 16

Correlations between meeting White benchmarks and other ethnic group benchmarks

		90 th percentile benchmark (White)	75 th percentile benchmark (White)	50 th percentile benchmark (White)	25 th percentile benchmark (White)
Corresponding African American benchmark	Partial correlation	.26**	.02	N/A	.09
	Significance	.00	.72	N/A	.13
Corresponding Latino benchmark	Partial correlation	.01	-.11	-.03	-.11
	Significance	.84	.06	.57	.06
Corresponding Asian benchmark	Partial correlation	.35**	.16*	-.04	.03
	Significance	.00	.01	.55	.63

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Achieving the benchmarks based on White students and the benchmarks based on Asian students were significantly related for the highest percentile groups (the 90th and 75th percentile groups) but not the lower percentile groups. Meeting the benchmarks based on White students were significantly related to reaching the benchmarks based on African American students for the 90th percentile group only and were not significantly related to reaching benchmarks based on Latino students at any level.

Benchmarks based on sex. Just as ethnic groups had different benchmark items that effectively distinguished between test-takers of that ethnic group, males and females may exhibit different benchmark items. The above-described process was repeated separately for males and females to generate sex-specific benchmarks. As before, all potential benchmarks are listed below with the selected benchmarks in bold.

Table 17

Benchmark items based on male and female students

Benchmark items based on male students		
90 th percentile	28²	Problem 3, Question 4⁴
75 th percentile	6, 32	Problem 4, Question 1²
50 th percentile	27¹	Problem 1, Question 3⁴
25 th percentile	.	.
Benchmark items based on female students		
90 th percentile	7, 17 ³ , 32, 41, 45 ²	Problem 5, Question 3⁴
75 th percentile	21 ⁴ , 28 ² , 35² , 39 ⁴	Problem 4, Question 1
50 th percentile	22	Problem 1, Question 3⁴
25 th percentile	.	.

¹ Memory subscale² Analytic subscale³ Creative subscale⁴ Practical subscale

As with the group overall, no items were found that could reliably distinguish between the 25th percentile group and those below the 25th percentile in either males or females. Males shared three out of seven benchmark items with the group as a whole (43%), all of which distinguished between the same percentile groups: Item 32 marked the 75th percentile; Problem 1, Question 3 marked the 50th percentile; and Problem 3, Question 4 marked the 90th percentile both in males and overall. Nearly twice as many suitable benchmarks were generated for females than were generated for males, 4 out of 13 (31%) of which were shared with the group as a whole: Item 22 marked the 50th percentile for both females and the overall group; Item 32 marked the 90th percentile in females but the 75th percentile overall; Item 35 marked the 75th percentile both in females and overall; and Problem 1, Question 3 marked the 50th percentile both for females and the overall group. The performance patterns of

females appeared to be not as well represented by the overall benchmarks as those of males.

Four benchmarks were shared by males and females, two of which distinguished between comparable groups and two of which functioned differently. Problem 4, Question 1 marked the 75th percentile in both males and females, and Problem 1, Question 3 marked the 50th percentile. Item 28 marked the 90th percentile for males but the 75th percentile for females, and Item 32 marked the 90th percentile in females but the 75th percentile in males. Out of the four benchmark items that were shared by males and females, the two that distinguished between the same percentile groups were both open-response items and the two that distinguished between different percentile groups were multiple-choice items.

In terms of the processes represented by benchmarks based on males and females, practical items again proved to be useful benchmarks. Out of the items that were associated with process subscales, 6 out of 13 (46%) were rated as practical items. For females, all benchmark items that were associated with a subscale came from the memory, analytic, or practical subscales whereas for males, benchmark items were associated with the analytic, creative, or practical subscales. As only one benchmark item each was associated with the memory subscale for females or the creative subscale for males, any seeming difference in the cognitive processes represented by male and female benchmarks cannot be considered reliable.

As before, binary variables were created to represent whether or not each individual met each of six additional benchmarks: the 90th percentile group based on males, the 90th percentile group based on females, the 75th percentile group based on

males, the 75th percentile group based on females, the 50th percentile group based on males, and the 50th percentile group based on females. For example, all participants who answered Item 28 correctly and received at least half credit on Problem 3, Question 4 (the 90th percentile benchmark items based on males) were coded as achieving the 90th percentile benchmark based on males. The relationships between achieving benchmarks based on males and those based on females with the overall benchmarks as well as comparisons between benchmarks based the two sexes appear in Table 18 below. All (phi) correlations controlled for total score to isolate any relationships between groups that were not related to their overall performance level.

Table 18

Correlations between sex group benchmarks

		Correlations between sex group benchmarks and overall benchmarks		
		90 th percentile benchmark (overall)	75 th percentile benchmark (overall)	50 th percentile benchmark (overall)
Corresponding male benchmark	Partial correlation	.29**	.61**	.40**
	Significance	.00	.00	.00
Corresponding female benchmark	Partial correlation	-.02	.13*	.37**
	Significance	.76	.03	.00
		Correlations between male and female benchmarks		
		90 th percentile benchmark (males)	75 th percentile benchmark (males)	50 th percentile benchmark (males)
Corresponding female benchmark	Partial correlation	.05	.25**	.36**
	Significance	.37	.00	.00

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Controlling for total score, meeting each male benchmark was significantly related to meeting the equivalent overall benchmark. Whereas meeting the 75th and 50th percentile benchmarks for females were significantly related to meeting the corresponding benchmarks for the overall group, meeting the 90th percentile

benchmark for females was not significantly related to meeting the 90th percentile benchmark overall. Similarly, meeting the 75th and 50th percentile benchmarks based on females were significantly related to meeting the corresponding male benchmarks but meeting the 90th percentile benchmark based on females was not.

Benchmarks by profile. Finally, different profile groups may exhibit different benchmarks because of their distinct patterns of strengths and weaknesses associated with process skills. The previously described benchmarking process was repeated separately for each of the profile groups, with benchmarks determined separately based on the direction of loading onto the factor. Thus, six different sets of benchmarks were generated, one for each of the six profiles of achievement. As before, all potential benchmarks are listed below with the selected benchmarks in bold. Note that because the item types varied in their usefulness as benchmarks, in many cases it was not possible to assign one multiple-choice item and one open-response item to be benchmarks for each profile group at each level.

Table 19

Benchmarks based on profile group

Benchmark items for Profile 1 (positive loading)		
90 th percentile	.	.
75 th percentile	20⁴	Problem 5, Question 2² Problem 5, Question 3 ⁴
50 th percentile	23, 35	Problem 2, Question 3²
25 th percentile	24⁴ , 28 ² , 37¹	.
Benchmark items for Profile 1 (negative loading)		
90 th percentile	9² , 17³	.
75 th percentile	1 ⁴ , 11 ² , 36	Problem 3, Question 3.2 ⁴ Problem 5, Question 2 ² Problem 5, Question 4
50 th percentile	.	.
25 th percentile	17³	Problem 1, Question 1.1²
Benchmark items for Profile 2 (positive loading)		
90 th percentile	1 ⁴ , 8 ³ , 13 ³ , 23 ² , 32, 35²	Problem 1, Question 3⁴ Problem 4, Question 3 ³ Problem 5, Question 3 ⁴
75 th percentile	6, 21 ⁴ , 27¹ , 40⁴	.
50 th percentile	25¹	Problem 2, Question 3
25 th percentile	5² , 41	Problem 2, Question 2²
Benchmark items for Profile 2 (negative loading)		
90 th percentile	7, 9 ² , 17 ³ , 18 ² , 23 ² , 31, 43 , 45 ²	Problem 1, Question 4 Problem 2, Question 3 Problem 3, Question 3.2⁴ Problem 5, Question 4
75 th percentile	1 ⁴ , 2 ² , 4 ³ , 6, 8 ³ , 14 ¹ , 19 ² , 22, 24 ⁴ , 27 ¹ , 29 ² , 32 , 35 ²	Problem 3, Question 4⁴
50 th percentile	10, 36	Problem 1, Question 1.2 ² Problem 1, Question 3 ⁴ Problem 2, Question 1 Problem 5, Question 1 ² Problem 5, Question 2 ²
25 th percentile	21⁴ , 35 ² , 42	.
Benchmark items for Profile 3 (positive loading)		
90 th percentile	6, 9 ² , 10 ⁴ , 14 ¹ , 16 ⁴ , 17 ³ , 21⁴ , 23 ² , 32, 35 ² , 41	Problem 3, Question 1.2 Problem 3, Question 4⁴
75 th percentile	20 ⁴ , 27 ¹	Problem 5, Question 3⁴
50 th percentile	11² , 22	.
25 th percentile	28²	Problem 1, Question 1.2 ² Problem 1, Question 3 ⁴ Problem 5, Question 4
Benchmark items for Profile 3 (negative loading)		
90 th percentile	.	Problem 3, Question 2.2² Problem 4, Question 2 Problem 5, Question 3⁴
75 th percentile	2 ² , 10 ⁴ , 24 ⁴ , 27 ¹ , 32, 37¹ , 40⁴ , 45 ²	.

50 th percentile	8 ³ , 34 ³ , 35 ²	Problem 4, Question 1 ² Problem 5, Question 4
25 th percentile	1 ⁴ , 6, 7, 19 ² , 21 ⁴ , 23 ² , 30 ³ , 31, 43	Problem 2, Question 1 Problem 2, Question 2 ² Problem 5, Question 2 ²

¹ Memory subscale

² Analytic subscale

³ Creative subscale

⁴ Practical subscale

For each group, benchmarks were very different from the benchmarks used to distinguish between performance levels overall. Profile 3 (positive loading) had the most similar benchmarks to the overall group with 8 out of 22 (36%) of benchmarks in common. Profile 1 (positive loading) shared 2 out of 10 (20%) of benchmarks, profile 1 (negative loading) shared 2 out of 11 (18%), profile 2 (positive loading) shared 4 out of 18 (22%), profile 2 (negative loading) shared 10 out of 36 (28%), and profile 3 (negative loading) shared 4 out of 28 (14%).

In comparing benchmarks of opposite profile types (e.g., profile 1 – positive loading versus profile 1 – negative loading), eight items were identified as possible benchmarks for both positive and negative loading versions of the same profile. However, items rarely served as benchmarks for the same level for both groups; the only items that served as benchmarks at the same level for opposite profile groups were items 23 and 27, which served as benchmarks for the 90th and 75th percentile groups (respectively) for all participants who exhibited profile 2, regardless of the direction of loading.

Out of benchmark items that corresponded to a subscale, most of them came from the analytic (44%) or practical subscales (30%). Across the board, items based on the processes participants were best at served as benchmarks either at the lowest levels

only or at every level. On the other hand, items based in the processes participants were worst at were useful benchmarks in distinguishing higher performers.

As before, binary variables were created to indicate whether each individual met the benchmarks based on each profile group. The relationships (calculated using phi correlations) between meeting the overall benchmarks and meeting the corresponding benchmark for each profile group (controlled for total score) appear in Table 20 below.

Table 20

Correlations between profile group benchmarks and overall benchmarks

		90 th percentile benchmark (overall)	75 th percentile benchmark (overall)	50 th percentile benchmark (overall)
Corresponding profile 1 (positive) benchmark	Partial correlation	N/A	.07	.01
	Significance	N/A	.24	.93
Corresponding profile 1 (negative) benchmark	Partial correlation	.02	-.05	N/A
	Significance	.77	.44	N/A
Corresponding profile 2 (positive) benchmark	Partial correlation	.00	.00	.05
	Significance	.94	.97	.42
Corresponding profile 2 (negative) benchmark	Partial correlation	.02	.45**	-.01
	Significance	.78	.00	.85
Corresponding profile 3 (positive) benchmark	Partial correlation	.39**	.06	.02
	Significance	.00	.29	.71
Corresponding profile 3 (negative) benchmark	Partial correlation	.27**	.01	-.02
	Significance	.00	.82	.70

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

There were few cases where meeting the benchmarks based on profile groups corresponded to meeting the benchmarks based on the overall group. Reaching the 90th percentile benchmark overall was significantly related to reaching the equivalent

benchmarks based on people who exhibited profile 3 (both positive and negative loadings), and reaching the 75th percentile benchmark overall was significantly related to reaching the equivalent benchmark based on people who exhibited profile 2 (negative loading). No other significant relationships between achieving benchmarks based on profile groups and overall benchmarks were found.

Discussion

This study yielded compelling answers to the four research questions it posed. First, the results indicate that it is possible to create a psychometrically-sound test that measures students' cognitive skills as well as content knowledge in the context of physics. Based on both classical test theory and the Rasch model, it appears as though the items on the augmented version of the AP Physics exam functioned well. The two sets of statistics indicated that, in general, people responded to questions in a way that was commensurate with their ability level. This is one of the most important qualities of an exam; if only low ability people answered difficult items correctly, this would indicate a serious problem in scoring or in the effectiveness of the overall test to accurately measure ability. Additionally, the data demonstrated that the items varied substantially in difficulty level, although both classical-test-theory statistics and the results of the Rasch analysis suggested that there were too many difficult items on the exam, especially with regard to the ability level of the test-takers. This is of particular importance to the Rasch model, which stipulates that the items closest to a person's ability level are the most effective at measuring the underlying construct. Given that there were few items at the ability level where most people were clustered, the degree to which these few items accurately distinguished between people of different

abilities is somewhat questionable. It is possible, however, that this particular version of the test would function more effectively if administered to the broad range of AP Physics students (exhibiting wider variability in achievement) nationwide. In addition, the goal of the test 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.

In terms of overall test functioning, statistics based both on classical test theory and the Rasch model suggest that this assessment was reliable in terms of its potential for consistently measuring individual ability and item difficulty. The high level of internal-consistency reliability suggests that the items measured the latent trait of physics knowledge despite the somewhat attenuated results due to the inclusion of a second latent dimension related to the measurement of cognitive processing skills. The factor-analytic results for the multiple-choice failed to yield a concise solution, suggesting that each item contributed uniquely to the overall assessment. This should not be surprising given that the items were developed with an eye toward both content coverage (5 topic areas) and cognitive processing skill (4 cognitive processing skills per topic), potentially resulting in 20 qualitatively different item types. So while content holds the test together in an internally consistent way, it is confounded with process information such that a small number of meaningful factors could not be extracted on the multiple-choice section. On the open-response section, however, a dominant factor emerged, suggesting that on this section the underlying construct of physics knowledge remained intact. Finally, that both content 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 item- and test-level findings indicate that the augmented version of the AP Physics Exam had good internal psychometric properties and, thus, it is possible to create a psychometrically-sound test in the domain of physics that incorporates cognitive process information. This is an important and necessary starting point; with the knowledge that the present assessment is valid and reliable, deeper questions about its potential to contribute to the measurement of physics knowledge over and above the current AP exam (i.e., incremental validity) can be addressed.

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 and his colleagues (2006) who identified six profiles of achievement among students on AP exams in psychology and statistics. More remarkable, the profiles exhibited by students on AP psychology, statistics, and physics exams were exactly the same. This 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 are independent of content. Thus, assessments that measure content alone are incomplete indicators of students' knowledge and 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 represented only 38% of students. Therefore, many students exhibited profiles that were not associated with strong memory or analytical skills; traditional tests that focus exclusively on these two process areas may fail to detect the relative strengths of these individuals.

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 easier than others (students found the memory subscale to be the easiest followed by the practical, creative, and analytic 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 discernable 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 the same proportions. This provides further support to the idea that these profiles are broadly generalizable; in addition to being unrelated to content, they are invariant across demographic groups. Identifying achievement profiles also yielded interesting information in terms of benchmarks. Specifically, this study found that items based on the cognitive processes participants were best at served as benchmarks either at the lowest levels only or at every level, whereas items based on the processes participants were worst at were useful benchmarks in distinguishing higher performers. That is, someone with strong creative skills and poor practical skills would be expected to perform consistent with their ability on creative items at every level, but that the demonstration of practical skills best distinguishes the highest performing students with this profile. It is important to stress that this finding demonstrates a relationship only; it does not follow that students who focus on improving the cognitive process areas where they are weak will achieve higher levels of performance. However, the finding that the highest levels of performance are most reliably characterized by strong performance in typically weak process areas calls attention to the usefulness of broad assessment tools in identifying truly exceptional students.

At the group level, an important goal of the present work was to investigate whether traditionally observed differences in achievement based on ethnicity and sex could be reduced using an assessment that incorporated a broad range of cognitive processes. Efforts were also taken to fully understand the nature of any persisting group differences by looking at what characterized high and low performance in various demographic groups. In many ways, the goal of decreasing group differences in achievement was met and certainly a greater understanding of the structure and progression of knowledge across demographic groups was realized.

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 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 analytic and memory skills would be only about half as much after the introduction of a

creative subscale. This is consistent with the findings of Stemler and his colleagues (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. Thus, it appears as though including a creative subscale benefits African American students regardless of content domain. One surprising finding was the reliably low test score difference between White students and African American students on the analytic 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 to 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 (demonstrated through high performance on past exams), the analytic skills of African American students selected into the AP program may be extremely high within their own ethnic group (e.g., they might be in the top 10% of their ethnic group in terms of analytical skills). 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 (e.g., students scoring above the 25th percentile on analytical reasoning) so the small ethnic differences observed on the analytic 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 analytic, creative, and practical subsections compared to White students, but noticeably larger differences on the memory subsection. This finding is consistent with previous research from Stemler and his colleagues (2006) on 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. The typically observed achievement gap between Latino students and White students ($d = -0.58$; Stemler et al., 2006) is somewhat reduced on both the analytic and creative subscales, although not on the practical subscale. This 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 students performed moderately, though significantly, worse than White students on the practical subscale. This represents a negative aspect of including the practical subscale, but the positive benefits of inclusion (e.g., its usefulness in providing benchmarks or its capacity for allowing students with profiles characterized by strong practical skills to demonstrate their content mastery) seem to more than outweigh this downside, particularly as the difference between White and Asian students' performance on this subscale does not translate into a reliable difference on the overall test.

Comparing benchmarks across ethnic groups revealed valuable information regarding the nature of differences in achievement based on ethnicity. Asian students showed the most similarity to White students: the benchmarks based on White students and the benchmarks based on Asian students were significantly related for the highest percentile groups (the 90th and 75th percentile groups) though not the

lower percentile groups. The benchmarks based on White students were also significantly related to the benchmarks based on African American students in the highest performing group. Although there was no relationship between White and Latino benchmarks, for other groups the structure of physics knowledge seems to be similar at high levels though not at lower levels. This sheds light on the nature of achievement differences across ethnic groups: it suggests that there are real differences in what people of different ethnic groups know at lower performance levels but that what people know at high levels is invariant (at least among White, African American, and Asian students). This is encouraging when considering that people who perform well on AP tests are more likely to pursue study that subject in college and beyond (Morgan & Maneckshana, 2000); thus, the people who are in the best positions to contribute to the educational and professional development of a field should not be expected to have different content knowledge or cognitive skills based on ethnicity. Because the present work can only speak to the structure and progression of knowledge within the context of physics, future research should replicate the process of identifying and comparing ethnicity-specific benchmarks in other subjects in order to examine whether these findings are unique to the domain of physics or are more broadly characteristic of how knowledge is acquired across ethnic groups.

Contrary to the expected findings, males and females differed in demonstrated ability on the creative and practical subscales but not on the memory or analytic subscales. This suggests that the typically observed achievement gap between males and females ($d = -0.37$; Stumpf & Stanley, 1996) is not the result of the type of

cognitive processes assessed by traditional measures. Indeed, augmenting the test resulted in somewhat higher sex differences on the test as a whole and on the practical subscale. The effect-size difference on the creative subscale was slightly less than what has been observed on typical assessments; this small decrease is outweighed by the larger increases in sex differences on the overall test and practical scales. These findings are counter to the study's predictions. Indeed, it was hypothesized that assessing a broader range of cognitive skills would allow females to better demonstrate their content knowledge as this procedure yielded such results in terms of ethnic differences in achievement. However, the present findings are not completely at odds with past research. Recall that researchers found that females perform worse on applied word problems but better on "pure mathematics" questions than males (O'Neill & McPeck, 1993). This distinction seems to correspond to the distinction the present research makes between practical and analytical items. In "pure mathematics" problems, students are primarily asked to evaluate an abstract problem (i.e., use analytical skills) whereas on an applied problem students are expected to use their knowledge of mathematics to address a concrete, real-world problem (i.e., use practical skills). Given this past research, that females are not benefited by including practical items is understandable. Certainly individual female students may benefit from this augmentation (particularly those who are strong in practical or creative skills), but it does not appear that assessing a broader range of cognitive skills is effective in reducing group differences in achievement based on sex.

Nonetheless, benchmarking revealed noteworthy differences in the structure of physics knowledge based on sex. Benchmarks based on males and those based on

females were highly related at every level except for at the 90th percentile level. Thus, whereas it appears that the structure of physics knowledge is different across most ethnic groups only at lower levels, what males know seems to differ from what females know only at the highest level. The inconsistent pattern of knowledge acquisition across sexes is somewhat concerning. One might hope that what characterizes mastery of a subject area is relatively invariant across demographic groups such that people at the highest levels of performance all know more or less the same thing. However, what the highest-performing males know is noticeably different than what the highest-performing females know. This has important implications for the nature of content knowledge at higher levels of education and in professional organizations since people who perform well on AP exams in a certain subject are more likely to major in and, as a result, pursue a career in that area (Morgan & Maneckshana, 2000). Thus, students who pursue physics in college and beyond might know appreciably different things based on their sex. Although the lack of shared knowledge at high levels appears problematic, it may have some positive effects. For example, bringing together accomplished people with diverse knowledge bases could yield interesting discussions in classrooms and innovative advances in professional fields. Furthermore, the number of people this finding affects should not be overstated. As the only noticeable differences between male and female benchmarks were observed for the 90th percentile group, only 10% of females demonstrated different physics knowledge than males. This corresponds to a small proportion of the population that might be expected to exhibit differences in physics knowledge based on their sex. As with ethnicity, future research would do well to examine whether this

pattern of knowledge acquisition based on sex is specific to physics or whether it applies more broadly to how males and females learn across all domains.

Overall, it appears that augmenting the AP Physics exam with creative and practical subscales effectively allowed underrepresented minorities to better express their content knowledge but was not an effective way to reduce the achievement gap observed between males and females. As this is the first study to explicitly investigate the effect of testing a range of cognitive processing skills on achievement differences based on sex, future research should attempt to determine whether these findings are generalizable or to what extent they are limited to physics. If these results are replicated across a range of content domains, an important goal of future research will be to explore other ways to allow females to better demonstrate their content mastery in this and other domains.

Future research might consider the effect of expanding the type of items used on tests on the achievement gap between males and females. This suggestion is based in part on this study's finding that out of the benchmarks shared by males and females, those that distinguished between the same percentile groups in both males and females were open-response items whereas those that distinguished between different percentile groups were multiple-choice items. That what females know is more comparable to what similarly performing males know on open-response items provides some evidence that males and females demonstrate their content knowledge differently based on the type of item. This is further supported by past research that has found reduced sex differences in performance on tests that emphasize writing (e.g., Stumpf & Stanley, 1996) and on those that use an open-response rather than a

multiple-choice format (Breland, Danos, Kahn, Kubota, & Sudlow, 1991; Bridgeman & Lewis, 1994; Mazzeo, Schmitt, & Bleistein, 1991), suggesting that expanding the range of item types might be a reasonable course of action for future research to pursue.

Limitations

There a number of noteworthy limitations to the present research. Among them, 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. This research had to infer what the achievement gap on the actual AP Physics exam might have been based on past research; although this inference is certainly reasonable, a stronger argument could have been made in favor of the augmented exam's capacity to reduce ethnic differences in achievement if this research had been able to directly compare scores on both versions of the AP Physics exam. In terms of the lack of data on college outcomes (e.g., admissions decisions or college majors), the present research was not able to assess the predictive power of this new measure relative to the traditional AP exam. As Sternberg and the Rainbow Project Collaborators (2006) found that augmenting the SAT with creative and practical subscales added to its predictive validity, one might expect similar results in the context of AP Physics, but this possibility cannot be confirmed with the current data.

Conclusion

This research demonstrates the usefulness and importance of developing tests that measure a broad range of cognitive skills on a number of fronts. Perhaps most superficially, practical skills were particularly useful in distinguishing among people of different performance levels. An important goal of psychometrics is to reliably and accurately measure the ability of test-takers and, as such, to the extent that benchmarks represent items that are particularly effective in meeting these goals, types of items that serve as strong benchmarks ought to be included in assessments. Furthermore, 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 et al., 2005; Geiser & 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. Finally, 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. Thus, in order to make the most valid inferences about students' content mastery, tests need to take these cognitive processing skills into account. Since 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; Sternberg & Spear-Swerling, 1996), on the SAT (Sternberg & The Rainbow Project Collaborators, 2006), and on AP exams (Stemler et al., 2006). As the extent of the benefits continues to expand across diverse contexts, future research should encourage the development of tests that measure both content and process information across additional domains. A challenge for future research will be to think creatively about how testing for broad cognitive skills can be incorporated into nontraditional domains. The effectiveness of assessing creative and practical skills in the context of physics (a domain that most people would regard as requiring primarily analytical skills) demonstrates the importance of reexamining assumptions regarding what cognitive processes are required by different subject areas and, therefore, what skills students should learn in the classroom and be held responsible for on exams. The concept of teaching and testing creative skills in physics, analytical skills in art, or practical skills in Spanish literature might seem peculiar, but has potential to allow individuals with strengths that do not meet current preconceptions of the skills a subject requires the chance to excel in new domains and make previously inconceivable contributions.

References

- ACT. (2008). *History of the ACT*. Retrieved March 28, 2008, from <http://www.act.org/aboutact/history.html>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Beaton, A., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy.
- Bejar, I. I. & Blew, E. O. (1981). Grade inflation and the validity of the Scholastic Aptitude Test. *American Educational Research Journal*, 18, 143-156.
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, Handbook I: Cognitive domain*. New York: Longmans Green.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bowman, M. L. (1989). Testing individual differences in Ancient China. *American Psychologist*, 44, 576-578.
- Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Sudlow, M. W. (1991). *A study of gender and performance on Advanced Placement History examinations* (College Board Report No. 91-4). New York: College Entrance Examination Board.

- Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37-50.
- Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT reasoning test scores add to high school grades: A straightforward approach*. (College Board Research Report No. 2004-04). New York: College Entrance Examination Board.
- Bridgeman, B., & Schmitt, A. (1997). Fairness issues in test development and administration. In W. W. Willingham & N. Cole, *Gender and fair assessment* (pp. 185-226). Mahwah, NJ: Lawrence Erlbaum Associates.
- Burton, E., & Burton, N. W. (1993). The effect of item screening on test scores and test characteristics. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 321-336). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Camara, W. & Echternacht, G. (2000). *The SAT I and high school grades: Utility in predicting success in college*. (Research Note, RN-10). New York: College Entrance Examination Board.
- 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). Bloom's taxonomy. Retrieved April 27, 2004, from <http://www.nwlink.com/~donclark/hrd/bloom.html>
- College Board. (2004). Exam scoring. Retrieved May 20, 2004, from

- <http://apcentral.collegeboard.com/article/0,3045,152-167-0-1994,00.html>
- College Board. (2007). *Advanced Placement report to the nation*. Retrieved December 15, 2007, from <http://apcentral.collegeboard.com>
- College Board. (2008a). *AP program*. Retrieved March 21, 2008, from <http://professionals.collegeboard.com/k-12/assessment/ap>
- College Board. (2008b). *Meet the SAT: What it means for your child*. Retrieved March 21, 2008, from <http://www.collegeboard.com/parents/tests/meettests/21295.html>
- 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*. (College Board Research Report No. 2002-09). New York: College Entrance Examination Board.
- Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude and Achievement Tests*. New York: College Entrance Examination Board.
- Dougherty, C., Mellor, L., & Jian, S. (2005). *The relationship between Advanced Placement and college graduation*. (NCEA Study Series 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, 176-206.
- Geiser, S., & Santelices, V. (2004). *The role of Advanced Placement and honors courses in college admissions*. Berkeley, CA: Center for Studies in Higher

- Education, University of California, Berkeley.
- Gollub, J. P., Bertenthal, M. W., Labov, J. B., & Curtis, P. C. (2002). *Learning and understanding. Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: National Academy Press.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York: Free Press.
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council.
- Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Jencks, C., & Phillips, M. (Eds.). (1998). *The black-white test score gap*. Washington, DC: Brookings Institution Press.
- Kelly, D. L., Mullis, I. V. S., Martin, M. O. (2000). *Profiles of Student Achievement in mathematics at the TIMSS international benchmarks: U.S. performance and standards in an international context*. Chestnut Hill, MA: TIMSS International Study Center.
- Klopfenstein, K. (2004). Advanced Placement: Do minorities have equal opportunity? *Economics of Education Review*, 23, 115-131.
- Labov, J. B. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. Washington, DC: National Research Council.
- Lane, S. (2004). 2004 NCME Presidential Address. Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23, 6-14.

- Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, Winter, 1-10.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practices*, 26, 3-16.
- Leighton, J. P., & Sternberg, R. J. (Eds.). (2004). *The nature of reasoning*. New York: Cambridge University Press.
- Leonard, D., & Jiang, J. (1999). Gender bias and the college prediction of the SATs: A cry of despair. *Research in Higher Education*, 40, 375-408.
- Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1991, April). *Do women perform better, relative to men, on constructed response tests or multiple-choice tests? Evidence from the Advanced Placement Examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.
- Morgan, R., & Maneckshana, B. (2000). *AP students in college: An investigation of their course-taking patterns and college majors*. (Statistical Report 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*. (Statistical Report 98-13). Princeton, NJ: Educational Testing Service.
- Mullis, I. V. S., Martin, M. O., Beaton, A., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Center for the

Study of Testing, Evaluation, and Educational Policy.

- Nettles, A. L., & Nettles, M. T. (Eds.). (1999). *Measuring up: Challenges minorities face in educational assessment*. Boston: Kluwer Academic.
- Noble, J. (2004). The effects of using ACT composite scores and high school averages on college admissions decisions for ethnic groups. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions*, (pp. 303-319). New York: RoutledgeFalmer.
- Nuñez, T. (1994). Street intelligence. In R.J. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 2, pp. 1045-1049). New York: Macmillan.
- Nuñez, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. New York: Cambridge University Press.
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum Associates.
- 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 groups*. (College Board Research Report No. 93-1). New York: College Entrance Examination Board.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.
- Ross, C. C. & Stanley, J. C. (1954). *Measurement in today's schools* (3rd ed). Upper

- Saddle River, NJ: Prentice Hall.
- Sawyer, R. L. (1985). *Using demographic information in predicting college freshman grades*. (ACT Research Report No. 87). Iowa City, IA: ACT, Inc.
- Shaw, E. (2005). *Researching the educational benefits of diversity*. (College Board Research Report No. 2005-4). New York: College Entrance Examination Board.
- Stanley, J. C., Benbow, C. P., Brody, L. E., Dauber, S., & Lupkowski, A. E. (1992). Gender differences on eighty-six nationally standardized aptitude and achievement tests. In N. Colangelo, S.G. Assouline, & D.L. Ambrosion (Eds.), *Talent development: Vol I. Proceedings from the 1991 Henry B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 42-65). Unionville, NY: Trillium Press.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, 52, 613-629.
- 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, 344-376.
- 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., & Grigorenko, E. L. (2000). *Teaching for successful intelligence: To increase student learning and achievement*. Arlington Heights, IL: Skylight Professional Development.
- Sternberg, R. J. & Spear-Swerling, L. (1996). *Teaching for thinking*. Washington, DC: American Psychological Association.
- Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence*, 34, 321-350.
- Sternberg, R. J., Torff, B., & Grigorenko, E. L. (1998). Teaching for successful intelligence raises school achievement. *Phi Delta Kappan*, 79, 667-669.
- Stumpf, H., & Stanley, J. C. (1996). Gender-related differences on the College Board's Advanced Placement and Achievement tests, 1982-1992. *Journal of Educational Psychology*, 88, 353-364.
- Stumpf, H., & Stanley, J. C. (1998). Stability and change in gender-related differences on the College Board Advanced Placement and Achievement tests. *Current Directions in Psychological Science*, 7, 192-196.
- United States Department of Education (2004). *Testing for results: Helping families, schools and communities understand and improve student achievement*. Retrieved March 29, 2008, from <http://www.ed.gov/nclb/accountability/ayp/testingforresults.html>

- Widaman, K. F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, 28, 263-311.
- Wilder, G. Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature*. (College Board Research Report No. 89-3). New York: College Entrance Examination Board.
- William, D. *Assessment for learning: Why no profile in US policy?* In J. Gardner (Ed.). *Assessment and learning* (pp. 169-183). London: SAGE Publications.
- Williams, B. (Ed.). (2004). *Closing the achievement gap: A vision for changing beliefs and practices*. Alexandria, VA: Association for Supervision & Curriculum Development.
- Willingham, W. W., & Cole, N. (Eds.) (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- 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 Press.
- Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.). *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289-301). New York: Routledge Falmer.
- Zwick, R. (2007). *College admission testing*. Alexandria, VA: National Association for College Admission Counseling.

APPENDIX A
Augmented AP Physics Exam

I. Multiple-Choice Section

1. The acceleration of gravity is about six times as much at the earth's surface as at the moon's surface. Rocks are dropped from the same height above the ground on the earth and the moon. Compared with the rock falling on the moon, the rock falling on the earth
 - a. falls about six times as far in the same time
 - b. achieves about six times the speed in the same time
 - c. takes about one sixth as long to reach the ground
 - d. is moving about six times as fast when it hits the ground
 - e. hits the ground at a time and speed that are different, but the difference depends on the masses of the rocks
2. A block of unknown material weighs 9.355 N in air and 8.865 N when suspended in water ($D_{\text{water}} = 1.0 \text{ g/cm}^3$). What is the volume of the block?
 - a. 50 cm^3
 - b. 490 cm^3
 - c. 500 cm^3
 - d. 5 cm^3
 - e. 50 m^3
3. Your friend tells you that a static shock from a doorknob can be as high as 15,000 Volts. He wants to know why a shock from an electrical outlet which is only 120 Volts is dangerous while a static shock is not. You answer correctly, saying:
 - a. "Static electricity is DC which is always safe, while the electrical outlet is AC which is not."
 - b. "It's not the voltage that is dangerous, but the amps."
 - c. "The electrical outlet conducts more electricity because it is transported along wires."
 - d. "The shock from the electrical outlet is more dangerous because it acts over a larger time, depositing more energy.?"
 - e. "The static shock has to travel through the air which reduces its power."
4. Assume that stars emit light and sound and that someone is able to detect both at some other planet in a different galaxy. How would the frequencies of the detected light and sound signals compare to those of the emitted ones?
 - a. The frequency of both will be higher.
 - b. Light will have higher frequency and sound lower frequency.
 - c. The frequencies do not change.
 - d. Light will have lower frequency and sound higher frequency.
 - e. The frequency of both will be lower.

5. A sodium surface with work function of 2.46eV is illuminated with light with energy of 4.14eV . What is the maximum kinetic energy of the ejected photoelectron?

- a. 0 eV
- b. 1.04 eV
- c. 1.68 eV
- d. 2.46 V
- e. 4.14 eV

6. If you walk 30 m at 2 m/s , and then run 30 m at 6 m/s , your average speed for the whole distance is

- a. 3 m/s
- b. 4 m/s
- c. 5 m/s
- d. 7.5 m/s
- e. 15 m/s

7. The flow speed above an airplane's wing is 20 m/s and the flow speed below the wing is 10 m/s . If the density of air is approximately 1.0 kg/m^3 , and the total wing surface area is 10 m^2 , determine the upward force on the airplane.

- a. 150 N
- b. 300 N
- c. 1500 N
- d. 3000 N
- e. 6000 N

8. Two small rocks, both negatively charged, are drifting in space. Their charges are so small that their gravitational attraction is greater than their electric repulsion -- so there is a net attraction. How can they be moved so that the forces exactly balance and there is no net force of attraction or repulsion?

- a. They are moved somewhat farther apart.
- b. They are moved much farther apart.
- c. They are put in orbit around each other.
- d. They are moved much closer together.
- e. There is no such possibility for these charges and masses.

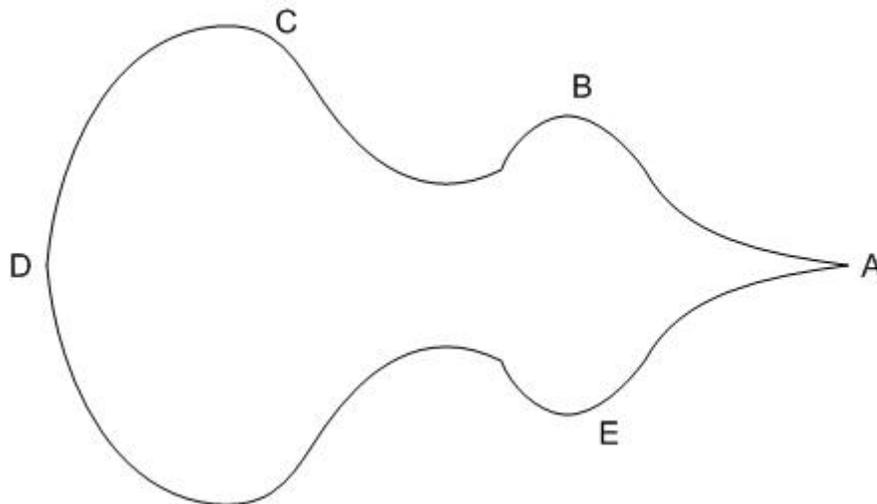
9. For a divergent thin lens, what is the correct sequence of signs for the following conditions?

- Object is in back of lens
- Image is in back of lens
- Image is upright
- Focal length
 - a. +, +, +, -
 - b. -, +, -, -
 - c. -, +, +, -.
 - d. -, +, +, +
 - e. -, -, +, -

10. You are a crash scene investigator. In analyzing a crash, you note a set of skid marks from one of the cars. The driver of this car asserts that he stopped as fast as he could.

- a. His statement is true because the skid marks indicate he pressed the brakes as hard as possible.
- b. His statement is false because if he had pressed harder on the brakes he would have stopped faster.
- c. His statement is true because brakes are designed to stop a car as fast as possible.
- d. His statement is false because a skidding car will travel farther than a car braked without skidding.
- e. There is not enough information to make the determination.

11. On a dry day, the object shown below gains electrostatic charges. Which area of the object will have the largest accumulation of charges?



- a. A
- b. B
- c. C
- d. D
- e. E

12. The wavelength of the third harmonic on a pipe with one closed end is $\lambda_3 = 1.33$ m. What is the wavelength of the next harmonic?

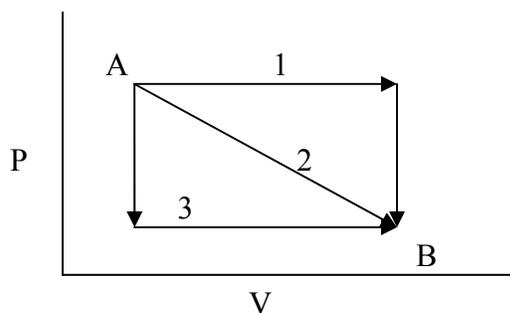
- a. 1.0 m
- b. 0.5 m
- c. 0.8 m.
- d. 1.6 m
- e. 4.0 m

13. You are required to measure the spring constant of a spring. You have the following items at your disposal: A stand to hang things from, a ruler, two identical known masses that can be hooked to the spring if desired, a stopwatch, and a length of string. Along with the spring, what is the smallest set of items you need to perform the measurement?

- a. The stand, the ruler, and the two masses.
- b. The stand, one mass, and the stopwatch.
- c. The stand and the ruler.
- d. The stand, both masses, and the stopwatch.
- e. The stand, the string, a mass and the ruler.

14. A Pascal is equivalent to which of the following in S.I. units?
- $\text{kg}/(\text{m}\cdot\text{s})$
 - $\text{kg}\cdot\text{m}/\text{s}$
 - $\text{kg}\cdot\text{m}^2/\text{s}$
 - $\text{kg}/(\text{m}^2\cdot\text{s}^2)$
 - $\text{kg}/(\text{m}\cdot\text{s}^2)$
15. A small object of mass 10.0 g carries a charge of $-20\ \mu\text{C}$. How should you tune an electric field generator (magnitude and direction) so that the mass is kept just suspended above the ground?
- 4.9 N/C, downward
 - 4900 N/C, downward
 - 4.9×10^6 N/C, downward
 - 0.049 N/C, upward
 - 4900 N/C, upward
16. Mariella needs a plain wall mirror to see her full height H . The local discount store has a mirror that is $(1/4)H$ in length. She places the mirror at eye level. How far back must she stand to see her full height in the mirror?
- A distance equal to half of her height.
 - A distance equal to her height.
 - A distance equal to twice her height.
 - A distance equal to three times her height.
 - No distance will work.
17. Suppose that Planck's constant was 1000 times larger than it actually is, what effect would this have on the wavelength of the photon emitted from the least energetic transition of the hydrogen Balmer series?
- The wavelength of the emitted photon would be 1000 times longer.
 - The wavelength of the emitted photon would be 10 longer.
 - The wavelength of the emitted photon would be 1000 shorter.
 - The wavelength of the emitted photon would be 10 shorter.
 - There would be no difference from what we currently experience.

18.



A system can move from state A to state B by each of the three paths shown in the figure above. Which of the following relationships is true?

- $w_2 > w_1$
- $\Delta U_1 > \Delta U_2$
- $q_1 = q_3$
- $\Delta U_1 = \Delta U_3$
- $w_3 = w_1$

19. If the frequency of the incident light on potassium metal is f , and the threshold frequency of potassium is t , and Plank's constant is h , then the kinetic energy of the ejected electron is given by:

- hf
- $h(f-t)$
- $h(t-f)$
- ht
- $h(f+t)$

20. A baseball is thrown horizontally from a window. It lands on the level ground 5 m from the base of the wall below the window. If the baseball is thrown horizontally from the same window at three times the original speed, how far from the wall will it land?

- 5 m
- $5(3)^{0.5}$ m
- 15 m
- $15(3)^{0.5}$ m
- 45 m

21. A child of mass, 12 kg, slides down a slide that is 3.0 m high and extends 3.0 m horizontally. The child comes to rest just as she gets to the bottom of the slide.

Approximately much internal energy was generated?

- 36 Joules
- 72 Joules
- 360 Joules
- 500 Joules
- 720 Joules

22. You are asked to build a capacitor with as large a capacitance as possible. You are given two small square metal plates, two large square metal plates, a sheet of thin paper, and a sheet of thick paper. Both pieces of paper are larger than the large metal plates. You choose:

- Large plates and the thin paper.
- Small plates and thin paper.
- Small plates and thick paper.
- Large Plates and thick paper.
- Large plates and thick and thin paper.

23. The isotope $^{210}_{82}\text{Pb}$ (lead-210) can emit, in quick succession, a negative electron, a gamma ray, another negative electron, and an alpha particle. What nucleus remains?

- $^{204}_{80}\text{Hg}$
- $^{205}_{82}\text{Pb}$
- $^{206}_{78}\text{Pt}$
- $^{206}_{82}\text{Pb}$
- $^{208}_{83}\text{Bi}$

24. 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.

- I only.
- II only.
- I and II
- I and III
- III only.

25. The unit of capacitance is dimensionally equivalent to which of the following

- J/C
- C/V
- V/C
- C/J
- J/V

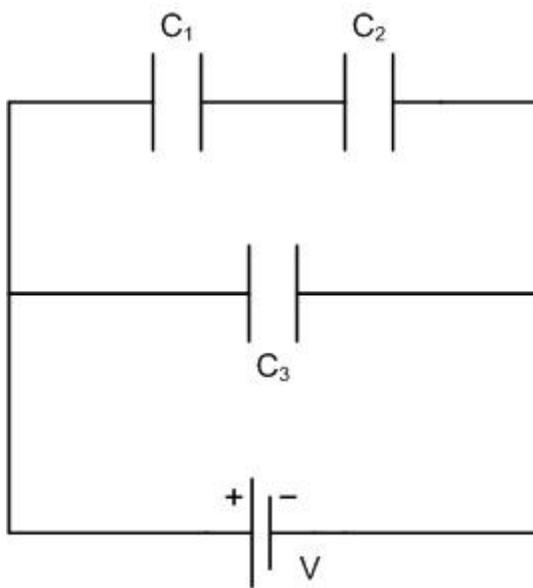
26. Computer screens have thin-film antireflection coatings. If the coating minimizes the reflection of green light ($\lambda = 510 \text{ nm}$), by how much should its thickness be changed to minimize the reflection of red light ($\lambda = 650 \text{ nm}$)?

- It should be increased 2.7%.
- It should be increased 27%.
- It does not have to be changed.
- It should be decreased 27%.
- It should be decreased 2.7%.

27. On a pressure-volume graph, an isobaric process is represented by what shape graph.

- a diagonal line
- a horizontal line
- a vertical line
- a parabola
- a hyperbola

28. Which of the statements is correct if $C_1 = C_2 = C_3$ in the following circuit



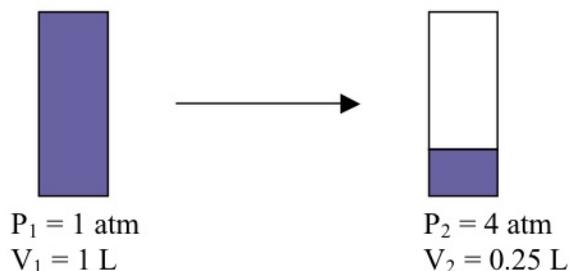
- $Q_1 = Q_2 = Q_3$
- $Q_1 = Q_2 > Q_3$
- $Q_1 = Q_2 < Q_3$
- $Q_1 < Q_2 < Q_3$
- $Q_1 > Q_2 = Q_3$

29. A mixture of two ideal gases is in thermal equilibrium at 500 K. Gas A has one-fourth the mass of a molecule of gas B. The rms speed of molecules of gas A is 600 m/s. what is the rms speed of molecules of gas B?

- 100 m/s
- 150 m/s
- 200 m/s
- 300 m/s
- 1200 m/s

30. The power output the Sun is approximately 4.0×10^{26} W. If the speed of light were half of its current value, how would the rate of mass loss of the Sun change relatively to its present rate?
- No change—Mass would be lost at the current rate.
 - Mass would be lost at twice the current rate.
 - Mass would be lost at four times the current rate.
 - Mass would be lost at one-half the current rate.
 - Mass would be lost at one-fourth the current rate.
31. Consider an object of mass m moving at constant speed v on a circle of radius r . The centripetal force needed to keep it in motion is
- proportional to the radius of the circle
 - proportional to its linear velocity
 - independent of the mass
 - inversely proportional to the radius of the circle
 - inversely proportional to the square of the speed
32. You have four 10 ohm resistors. What value of equivalent resistance can you not obtain by combining all four of these resistors?
- 2.5 ohms
 - 10 ohms
 - 13.3 ohms
 - 20 ohms
 - 40 ohms
33. Ammar holds a shiny teaspoon at arms length and looks into the back of the spoon. He observes that his image is upright. He then turns the teaspoon around so that he is looking into the bowl of the spoon. The image that he sees is:
- Upright and larger.
 - Upright and smaller.
 - Upside down and larger.
 - Upside down and smaller.
 - Upside down and the same size.
34. You are in a space station orbiting the earth. You have two wooden spheres that are outwardly identical. One of them has a lead core. What could you do to determine which one has the lead core? It's impossible you need some sort of measuring device.
- The solid wood ball will be easier to shake.
 - The one with the lead core will wobble if you spin it.
 - If you release them, the solid wood one will drift more quickly toward the wall of the space station.
 - If you release them, the solid wood ball will drift noticeably toward the lead cored ball.

35.



If a gas undergoes the changes shown in the figure above, by what factor does the internal energy of the gas change?

- Internal energy increases by a factor of 4
- Internal energy decreases by a factor of 4
- Internal energy increases by a factor of 16
- Internal energy decreases by a factor of 2
- Internal energy remains the same

36. As a block slides down a frictionless inclined plane, of angle θ , a graph of work done by the gravitational force versus distance down the plane is plotted. The shape of the graph will be

- a horizontal line at mg
- a horizontal line at $mg\sin\theta$
- a diagonal line with slope mg
- a diagonal line with slope $mg\sin\theta$
- a diagonal line with slope $mg/\sin\theta$

37. An ideal gas has its temperature changed from 300 K to 600 K. By what factor does the root mean square speed of the molecules change?

- 4
- 2
- $\sqrt{2}$
- 0.5
- 10

38. Imagine that the speed of light were smaller than the speed of sound. What would happen?

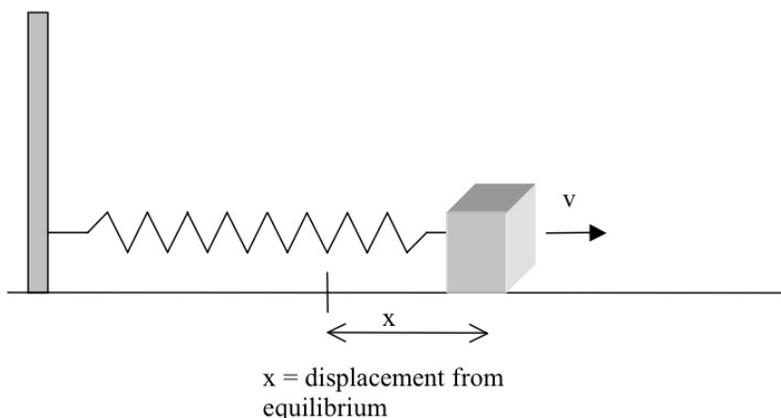
- You would hear the thunder after you see the lightning
- You would see the lightning after you hear the thunder
- You would hear the thunder and see the lightning at the same time
- You would never see the lightning
- You would never hear the thunder

39. When you drive a car at constant speed around a circular track, the net force on you
- acts towards the inside of the track and is proportional to the square of your speed
 - acts towards the inside of the track and is proportional to your speed
 - is zero
 - acts forward and is proportional to your speed
 - acts toward the outside of the track and is proportional to the square of your speed
40. A woman stands in front of a makeup mirror and sees her face at twice its actual size and right-side up. If she is 30.0 cm in front of the mirror, what is the mirror's focal length?
- 20.0 cm
 - 30.0 cm
 - 40.0 cm
 - 60.0 cm
 - 80.0 cm
41. You have two identical spring scales that will measure the weight of an object up to 500 grams. You have an object you wish to weigh but it weighs more than 500 grams but less than 1kg. What should you do?
- Put the springs in series, then add the two readings.
 - Put the object on a flat surface and pull it with one of the springs and read off the value.
 - Put the springs in parallel and then double the reading of one scale.
 - Put the springs in parallel and then divide the reading of one scale by 2.
 - Put the springs in parallel and then add the inverses of each spring's reading.
42. Two physics students Thelma and Louise are investigating the conservation of momentum. They have two balls of equal mass 5 kg. If they each launch a ball at 2 m/s in opposite directions what will be the total momentum of the system?
- 20 kg m/s
 - 20 kg m/s
 - zero
 - 10 kg m/s
 - 10 kg m/s
43. The law that states that “ the induced emf in a conducting loop equals the rate of change of magnetic flux through the circuit” is credited to whom?
- Ampere
 - Newton
 - Volta
 - Ohm
 - Faraday

44. Imagine a world in which electrons are positive and protons are negative. In that Universe

- opposite charges will repel and equal charges attract.
- opposite charges will attract and equal charges repel.
- all charges will attract.
- all charges will repel.
- It cannot be predicted.

45.



A block weighing 10 N is oscillating on a frictionless surface as shown in the figure above. If the period of motion of the block is 2.5 s, find the spring constant.

- 63 N/m
- 6.3 N
- 50.1 N/m
- 4.0 N/m
- 31.6 N/m

II. Open-Response Section

PROBLEM 1

Introductory Information (*relevant to all subsequent questions for problem 1*)

You are a member of the first group from Earth to land on the surface of Europa, a satellite of Jupiter. It is known from accurate telescopic observations that Europa is a sphere of radius 1570 km; it has no atmosphere.

Problem 1, Question 1

Inside the spacecraft you place a 0.500 kg mass on an electronic balance. It weighs 0.66 N.

- How large is the acceleration of gravity on Europa?
- Calculate the mass of Europa.

Problem 1, Question 2

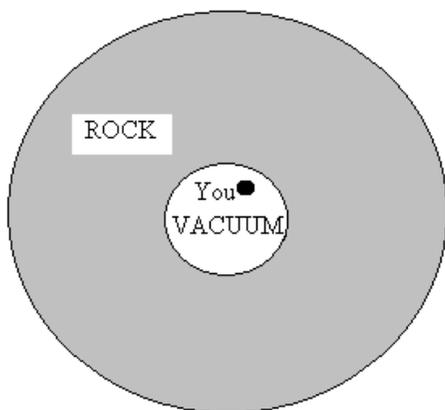
Upon landing, you want to be sure you can protect Europa from outsiders. You practice shooting cannonballs off a hill. Your first shot arcs downward and hits the ground of Europa with a higher velocity. On your second shot, you fire the cannonball fast enough to go into circular orbit. The velocity of this shot does not increase. Explain why the velocity of the first shot increases, while the velocity of the second shot remains the same.

Problem 1, Question 3

On Earth, you can throw a baseball 30 m straight upward. How high can you throw a baseball on Europa?

Problem 1, Question 4

Millions of years ago, an advanced civilization hollowed out the center of Europa. They built a strong spherical shell centered on the center of Europa, whose radius is 100 km. Inside it is a vacuum. Suppose you are floating in the vacuum 1.0 km inside this shell. Why is your net acceleration zero? Base your argument on Newton's Law of Gravitation and state your reasoning. Assume that the mass distribution of Europa is spherically symmetric.

**PROBLEM 2****Introductory Information** *(relevant to all subsequent questions for problem 2)*

A fiber optics cable has a core with an index of refraction of 1.60 and an outer layer called the cladding with an index of refraction of 1.20. The cladding is surrounded by a third protective layer.

Problem 2, Question 1

What is the speed of light in the core?

Problem 2, Question 2

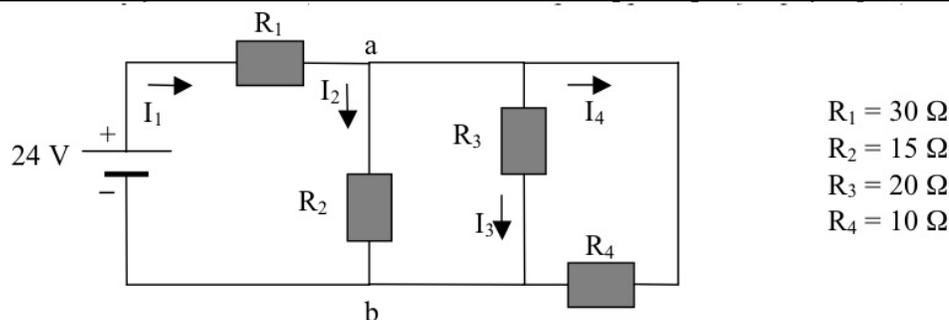
Determine the critical angle of a beam of light that is traveling in the core and hits the cladding.

Problem 2, Question 3

Explain why a cladding is necessary in a fiber optics cable.

Problem 2, Question 4

In many medical applications of fiber optics it is necessary to bend the cable around corners. Explain whether it is necessary to have a lower or higher critical angle between the core and the cladding for the light beam to remain within the core.

PROBLEM 3**Introductory Information** (*relevant to all subsequent questions for problem 3*)

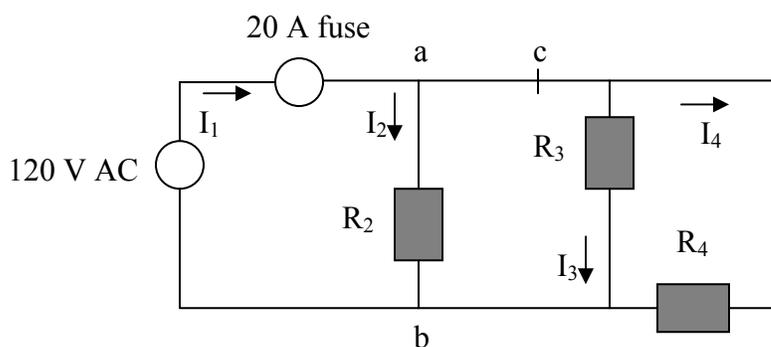
A circuit is constructed using a 24-V battery and four resistors positioned as shown in the figure above.

Problem 3, Question 1

- (i) Which of the four currents is the largest? Explain your reasoning.
- (ii) Which of the four currents is the smallest? Explain your reasoning.

Problem 3, Question 2

- (i) Find the equivalent resistance of the circuit.
- (ii) Determine the current through I_1 and the voltage drop between points a and b.

Problem 3, Question 3


$R_1 = 1000$ Watt microwave $R_2 = 1250$ Watt iron $R_3 = 600$ Watt blender

A circuit similar to the original one is used to wire the kitchen in a house. The power source is now the usual 120 V AC found in most American homes. In place of resistors 2, 3 and 4 are appliances: a 1000-watt microwave, a 1250-watt iron, and a 600-watt blender. There is a 20 Amp fuse in place of resistor 1. The new circuit is shown above.

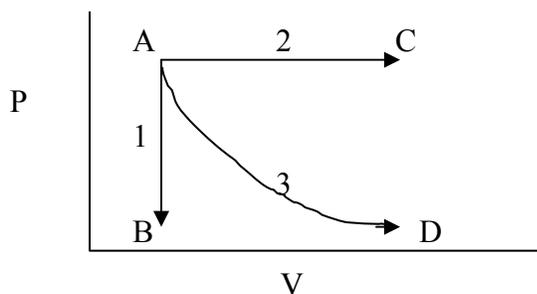
- What is the voltage drop between points a and b?
- If all three appliances are turned on at the same time, will the fuse burn out?

Problem 3, Question 4

If the fuse is moved to point c in the circuit, will the total current change? Will the fuse burn out? Explain why this placement of the fuse is not advisable.

PROBLEM 4

Introductory Information (*relevant to all subsequent questions for problem 4*)



- $P = 2.5$ atm, $V = 15$ L
- $P = 0.5$ atm, $V = 15$ L
- $P = 2.5$ atm, $V = 75$ L
- $P = 0.5$ atm, $V = 75$ L

Three samples of an ideal gas (1.5 moles each) are confined under identical conditions so that they are all initially in state A. Gas 1 undergoes a process that takes it to state B, gas 2 undergoes a process that takes it to state C, and gas 3 undergoes a process that takes it to state D.

Problem 4, Question 1

Complete the table that identifies the process undergone by each gas and give the correct sign for work.

Gas	Step	Process	Sign of Work (+, -, 0)
1	A → B		
2	A → C	Isobaric Expansion	
3	A → D	Isothermal Expansion	negative

Problem 4, Question 2

Calculate the work for the process undergone by gas 2. Calculate the heat and work for the process undergone by gas 3. The appropriate value for R is 0.0821 L atm/K mol.

Problem 4, Question 3

All the important thermal process besides adiabatic were represented in the PV curves. Think of an example of a situation that would result in an adiabatic process. What would the adiabatic process look like on a PV curve?

Problem 4, Question 4

You want to use one or more of the processes as part of a heat engine. A heat engine takes in energy by heat and partially converts it to other forms. Following this definition, which step(s) might you be able to use as a heat engine? Which step(s) would you not be able to use as a heat engine? Explain your reasoning.

PROBLEM 5

Introductory Information (*relevant to all subsequent questions for problem 5*)

A research lab is attempting to develop a metal alloy with a low work function to market it for practical applications of the photoelectric effect. The following questions pertain to the development of this alloy.

Problem 5, Question 1

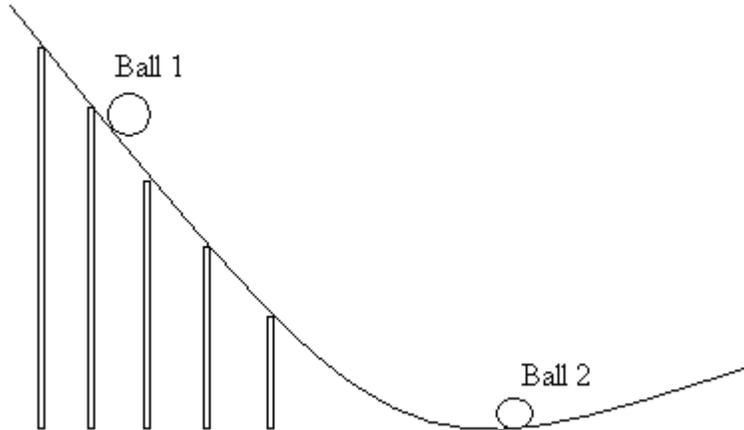
A work function of 2.0 e.v. is desired. Photons of what frequency would be required to remove electrons from this alloy?

Problem 5, Question 2

What wavelength of light would this frequency correspond to?

Problem 5, Question 3

The photoelectric effect can be modeled with an incline track and small balls, as shown below. Ball 1 is released at various heights on the incline track, causing Ball 2 to move different distances. What particle does Ball 1 represent? What particle does Ball 2 represent? What does releasing Ball 1 at various heights represent?



Problem 5, Question 4

The marketing department of this research lab wants to use this new alloy for semiconductors. In semiconductors, incoming light can move electrons from a low energy valence band to a higher-energy conduction band. The electrons in the conduction band can then be turned into an electric current. Why might the new alloy make a good semiconductor?

APPENDIX B

Answer Key and Scoring Guideline for the Augmented AP Physics Exam

I. Multiple-Choice Section

1. B	24. D
2. A	25. B
3. D	26. C
4. E	27. B
5. C	28. C
6. A	29. D
7. C	30. C
8. E	31. D
9. C	32. D
10. D	33. D
11. A	34. C
12. C	35. E
13. B	36. D
14. E	37. C
15. B	38. B
16. E	39. A
17. A	40. D
18. D	41. C
19. B	42. C
20. C	43. E
21. C	44. B
22. A	45. B
23. D	

II. Open-Response Section**Problem 1**

<u>Question 1</u>	<u>Points</u>
(i) $W = mg$ $0.66 = 0.5g$, so $g = 1.32 = 1.3 \text{ m/s}^2$	2
(ii) $F = GMm/r^2$ $0.66 = 6.67 \times 10^{-11} M \times 0.5 / (1.57 \times 10^6)^2$ $M = 4.9 \times 10^{22} \text{ kg}$ Write both formulas Arithmetic and units correct	2 1 1

Question 2	Points
The first shot arcs downwards, such that the acceleration of gravity on Europa comes into play.	1
Europa's gravitational force will act on the cannonball as it moves downward, increasing the cannonball's velocity.	2
The second shot is fired fast enough that it immediately goes into circular motion. Therefore, Europa's gravitational force will always be perpendicular to the second cannonball's velocity.	2
Since the acceleration and velocity are perpendicular, the velocity of the second shot will remain the same.	1

Question 3	Points
Throwing upward: $v^2 = v_0^2 + 2ax$	1
Substitution get $v_0 = 24.2$ m/s	2
On Europa, use same formula with $g = 1.32$, $x = 220$ m	2

Question 4	Points
Newton's Law is basically the same as Coulomb's Law.	3
We know that inside a sphere with charge evenly distributed on it, there is no electric field. So there is no force on a mass inside Europa.	3
<u>or</u>	
The rock nearby attracts me more than the rock far away because of the $1/r^2$ factor.	3
But there is much more mass of rock on the far side. This could make the acceleration towards the center, or exactly balance the force the opposite direction.	3

Problem 2

Question 1	Points
$v=c/n$	1
$=3.0 \times 10^8 \text{m/s}/1.6$	1
$=1.88 \times 10^8 \text{m/s}$	1
Question 2	Points
$\theta = \sin^{-1} n_2/n_1$	1
$= \sin^{-1} 1.2/1.6$	1
$= 49^\circ$	1

Question 3	Points
There must be a material with a lower index of refraction for total internal reflection to occur.	4
<u>or</u>	
Cables are buried so air cannot be used.	4
<u>or</u>	
There needs to be a protective outer layer.	4

Question 4	Points
Lower critical angle, since with a bent fiber optics cable the light beam will hit the interface with the cladding at many different angles depending on whether it hits the outer radius or inner radius.	4

Problem 3

Question 1	Points
I_1 is the largest.	1
All of the current goes through R_1 so for correct $I_1 = I_2 + I_3 + I_4$, therefore it must be the largest.	1
I_3 is the smallest.	1
The resistors R_2 , R_3 , and R_4 are in parallel therefore the voltage drop across them is the same. Since R_3 is the largest resistor of the three, it must carry the smallest current.	1

Question 2	Points
$1/R_p = 1/R_2 + 1/R_3 + 1/R_4$	1 point for correct R_p
$1/R_p = 1/15 \Omega + 1/20 \Omega + 1/10 \Omega = 13/60 \Omega \quad R_p = 4.6 \Omega$	
$R_s = R_1 + R_p = 30 \Omega + 4.6 \Omega = 34.6 \Omega$	1 point for correct R_s
$I_1 = 24 \text{ V} / 34.6 \Omega = 0.69 \text{ A}$	1 point for correct current
$V_{ab} = 24 \text{ V} - (30 \Omega \times 0.69 \text{ A}) = 3.3 \text{ V}$	1 point for correct voltage drop across R_1 1 point for correct value for V_{ab}

Question 3 Points
 The voltage drop from a to b is the voltage of the power source, 120 V. 1

$I_{\text{total}} = 23.7 \text{ A}$ 1

The current is more that the fuse will allow so it will burn out. 1

Question 4 Points
 The current will be the same. 1

The fuse will not burn out since the current through it is now 15.4 A. 1

The current through the circuit is still 23.7 A, but the fuse is no longer protecting the circuit. 1

Problem 4

Question 1 Points

Gas	Step	Process	Sign of Work (+, -, 0)
1	A → B	Isochoric (V constant)	zero
	A → C	Isobaric expansion	negative
3	A → D	Isothermal Expansion	negative

1 point for each correct response

Question 2 Points

For gas 2: $W = -P\Delta V$
 $W = -2.5 \text{ atm} (75\text{L} - 15\text{L})$
 $W = -150 \text{ L atm}$ 2
 (1 for answer, 1 for sign)

For gas 3: $T = PV/nR$
 $T = \frac{2.5 \text{ atm} \times 15 \text{ L}}{1.5 \text{ mol} \times 0.0821 \text{ L atm/K mol}}$
 $T = 304.5 \text{ K}$ 1

$W = -nRT \ln(V_2/V_1)$
 $W = -1.5 \text{ mol} \times 0.0821 \frac{\text{L atm}}{\text{K mol}} \times 304.5 \text{ K}$
 $\times \ln(75\text{L}/15\text{L})$
 $W = -60.4 \text{ L atm}$ or -6110 J 1

Since process is isothermal $\Delta U = 0$
 Therefore $Q = -W$
 $Q = 60.4 \text{ L atm}$ or -6110 J 1

<u>Question 3</u>	<u>Points</u>
Any example where no heat flows into or out of the system (or a reasonable approximation of no heat flowing into or out of the system) is an acceptable answer. Examples: compression of a gas in an insulated cylinder, fluid flow through a nozzle, process inside any insulated wall or system, atmosphere temperature as a function of height	2
The shape must be similar to that of an isotherm	1
<u>or</u>	
The volume decreases, as the pressure increases.	1
NOT identical to an isotherm because the temperature changes in an adiabatic process	1
<u>or</u>	
Adiabatic process curve should be steeper than isotherm	1
<i>Note.</i> If student draws a PV curve, instead of writing what the curve looks like:	
<ul style="list-style-type: none"> - determine if the curve drawn follows the trend of the volume decreasing, as the pressure increases for awarding the first point - comparison of adiabatic curve is made to an isotherm on the PV-curve, showing that the adiabatic curve is different OR steeper than an isotherms curve for second point 	

<u>Question 4</u>	<u>Points</u>
Steps A \rightarrow D (gas 3) are an isothermal expansion, such that temperature does not change so it could not be used as a heat engine.	2
Steps A \rightarrow C (gas 2) has a temperature rise by Charles's Law. By the definition given, it also could not be used as a heat engine since the temperature increases, or heat is given out by the step.	2
Steps A \rightarrow B (gas 1) has a temperature decrease by Gay-Lussac's Law. By the definition given, this step COULD be a heat engine since the accurate temperature decreases, or heat is taken in by the step.	2

Problem 5

<u>Question 1</u>	<u>Points</u>
$E=hf$	
$f=E/h=2.0 \text{ e.v.}/4.14 \times 10^{-15} \text{ e.v.-s}=4.83 \times 10^{14} \text{ hz}$	4
<u>Question 2</u>	<u>Points</u>
$\text{Wavelength}=c/f=3.0 \times 10^8/4.83 \times 10^{14}=6.2 \times 10^{-7} \text{ m}=620 \text{ nm}$	3
<u>Question 3</u>	<u>Points</u>
Ball 1 is like the photons hitting a material.	1
Ball 2 is like the electrons released from a material when photons hit the material.	1
The various heights Ball 1 is released at represent photons of different energies hitting the material	2
<u>Question 4</u>	<u>Points</u>
With the low work function, only low energy light is needed to move electrons from the low energy valence band to the higher-energy conduction band.	2
The more rapidly and easily electrons can be moved the greater the electric current that can be created, such that the low work function also allows for a greater electric current.	2

APPENDIX C
Item Difficulties and Discrimination Values

I. Multiple-Choice Items

Item	Difficulty	Discrimination (based on section)	Discrimination (based on test)
1	0.35	.26	.20
2	0.37	.06	.07
3	0.14	.07	.13
4	0.24	.20	.11
5	0.80	.22	.19
6	0.51	.33	.27
7	0.37	.23	.17
8	0.23	.17	.23
9	0.23	.13	.03
10	0.36	.27	.14
11	0.60	.05	.00
12	0.17	.09	.14
13	0.31	.08	.04
14	0.46	.35	.30
15	0.26	.21	.22
16	0.29	.22	.16
17	0.52	.22	.23
18	0.30	.06	.00
19	0.50	.35	.30
20	0.75	.33	.30
21	0.55	.35	.37
22	0.56	.35	.15
23	0.49	.29	.20
24	0.63	.31	.20
25	0.53	.38	.32
26	0.15	-.03	-.01
27	0.54	.45	.38
28	0.54	.25	.23
29	0.24	.40	.30
30	0.15	.14	.15
31	0.60	.33	.34
32	0.33	.49	.42
33	0.46	.12	.00
34	0.21	.26	.13
35	0.48	.34	.30
36	0.45	.17	.04
37	0.41	.41	.24

38	0.85	.26	.23
39	0.49	.38	.31
40	0.48	.07	-.05
41	0.41	.32	.23
42	0.78	.26	.19
43	0.71	.12	.17
44	0.78	.24	.16
45	0.35	.30	.25

II. Open-Response Items

Item	Mean score	Discrimination (based on section)	Discrimination (based on test)
Prob1Q1.1	1.82	.30	.29
Prob1Q1.2	1.74	.36	.35
Prob1Q2	1.31	.24	.27
Prob1Q3	2.89	.43	.45
Prob1Q4	0.52	.18	.19
Prob2Q1	2.35	.37	.37
Prob2Q2	2.64	.16	.18
Prob2Q3	2.83	.24	.27
Prob2Q4	1.41	.13	.09
Prob3Q1.1	0.87	.23	.23
Prob3Q1.2	0.79	.29	.32
Prob3Q2.1	0.63	.27	.21
Prob3Q2.2	0.31	.33	.34
Prob3Q3.1	0.19	.24	.27
Prob3Q3.2	0.88	.40	.41
Prob3Q4	1.14	.38	.34
Prob4Q1	1.49	.34	.36
Prob4Q2	0.76	.18	.11
Prob4Q3	0.97	.22	.20
Prob4Q4	0.92	.18	.16
Prob5Q1	2.36	.49	.50
Prob5Q2	1.74	.40	.41
Prob5Q3	2.34	.29	.27
Prob5Q4	0.87	.27	.25