Available online at www.sciencedirect.com
SCIENCE DIRECT®
Intelligence 34 (2006) 321–350
INTELLIGENCE

The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills

Robert J. Sternberg
The Rainbow Project Collaborators [1]

Tufts University, Office of the Dean of Arts and Sciences, Ballou Hall 3rd Floor, Medford MA 02155, United States

Received 17 July 2003; received in revised form 29 November 2005; accepted 3 January 2006
Available online 24 February 2006

Abstract

This article describes the formulation and execution of the Rainbow Project, Phase I, funded by the College Board. Past data suggest that the SAT is a good predictor of performance in college. But in terms of the amount of variance explained by the SAT, there is room for improvement, as there would be for virtually any single test battery. Phase I of the Rainbow Project, described here, uses Sternberg's triarchic theory of successful intelligence as a basis to provide a supplementary assessment of analytical skills, as well as tests of practical and creative skills, to augment the SAT in predicting college performance. This assessment is delivered through a modification of the Sternberg Triarchic Abilities Test (STAT) and the development of new assessment devices. Results from Phase I of the Rainbow Project support the construct validity of the theory of successful intelligence and suggest its potential for use in college admissions as an enhancement to the SAT. In particular, the results indicated that the triarchically based Rainbow measures enhanced predictive validity for college GPA relative to high school grade point average (GPA) and the SAT and also reduced ethnic group differences. The data suggest that measures such as these potentially could increase diversity and equity in the admissions process.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Intelligence; Rainbow Project; Analytical ability; Creative ability; Practical ability

1. Introduction

Standardized tests are frequently used in the United States and abroad as a basis for making high-stakes decisions about educational opportunities, placements, and diagnoses. One of the most widely used tests for these purposes is the SAT. Many colleges and universities in the U.S. use the SAT, usually taken during the high school years, as a predictor of success in college.

The SAT I is a 3-h examination that measures verbal comprehension and mathematical thinking skills (also referred to as reasoning abilities); new in recent years is an added writing component. A wide variety of studies

E-mail address: robert.sternberg@tufts.edu.
[1] Rainbow Project Collaborators: Damian Birney, Brent Bridgeman, Anna Cianciolo, Wayne Camara, Michael Drebot, Sarah Duman, Richard Duran, Howard Everson, Ann Ewing, Edward Friedman, Elena L. Grigorenko, Diane Halpern, P. J. Henry, Charles Huffman, Linda Jarvin, Smaragda Kazi, Donna Macomber, Laura Maitland, Jack McArdle, Carol Rashotte, Jerry Rudmann, Amy Schmidt, Karen Schmidt, Brent Slife, Mary Spilis, Steven Stemler, Robert J. Sternberg, Carlos Torre, and Richard Wagner. Further details on the collaborators are available from the authors.

0160-2896/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2006.01.002

have shown the usefulness of the SAT as a predictor of college success (Bridgeman, McCamley-Jenkins, & Ervin, 2000; Ramist, Lewis, & McCamley-Jenkins, 1994; Willingham, Lewis, Morgan, & Ramist, 1990). Each SAT II is a 1-h subject-specific test that measures achievement in designated areas such as mathematics, foreign languages, various sciences, and so forth.

A recent meta-analysis of the predictive validity of the SAT, encompassing roughly 3000 studies and more than one million students, suggested that the SAT is a valid predictor of early-college academic performance (as measured by first-year grade point average [GPA]), with validity coefficients generally in the range of .44 to .62 (Hezlett et al., 2001). The validity coefficients for later-college performance were somewhat lower but still substantial—generally ranging from the mid .30s to the mid .40s. Ramist et al. (1994) found that the validity of the SAT I at 38 colleges was better than high school GPA for predicting one specified course grade, but that high school GPA was a better predictor of overall first-year GPA. The correlations (corrected for restriction of range and criterion unreliability) with freshman GPA were .60 for SAT-Verbal (SAT-V), .62 for SAT-Math (SAT-M), .65 for SAT-Combined (SAT-C), and .69 for high school GPA. The corrected multiple correlation of high school GPA and SAT-C with freshman grades was .75. SAT-V and SAT-M predicted differentially for different courses. The difference favoring the SAT-V scores was greatest (~30%) for various types of English courses and history. The differences favoring the SAT-M scores were greatest (~35%) for mathematics and physical sciences/engineering courses. Correlations for females were generally higher than for males. Correlations also differed somewhat for different ethnic groups: the predictive effectiveness of the SAT-C varied from the highest (.64) for White students to the lowest (.50) for Native American students, with Asian American (.63), Black (.62), and Hispanic (.53) students taking intermediate positions in the order specified here.

Kobrin, Camara, and Milewski (2002) examined the validity of the SAT for college admission decisions in California and elsewhere in the United States. They found that, in California, SAT I and SAT II both showed moderate correlations with family income (in the range of .25 to .55 for SAT I and in the range of .21 to .35 for SAT II) and parental education (in the range of .28 to .58 for SAT I and in the range of .27 to .40 for SAT II). These findings indicate that SAT scores may be a function, in part, of social class. Predictive effectiveness of the SAT was similar for different ethnic groups; however, there were important mean differences and differences in changes in score across time (see also Bridgeman, Burton, & Cline, 2001). The group differences are reflected by the number of standard deviations (SD) away from the White students' mean each group scored. When all SAT scores were aggregated (i.e., when both SAT-I and SAT-II scores were considered), in comparison with White students on average, African American students scored about one full SD lower, Latino students scored 0.9 SD lower, and Native Americans scored about half a SD lower. Asian students demonstrated slightly lower scores (by .2 of SD) than did White students for the aggregated score. In particular, they scored higher than White students by about .03 (SAT I) to .07 (SAT II) SDs on the math tests, but about a third (SAT I) to half a (SAT II) SD lower on the verbal/writing tests.

All together, these results suggest good predictive validity for the SAT for freshman college performance. But as is always the case for any single test or type of test, there is room for improvement. The prediction of tests of "general" ability typically can be improved upon, and there is evidence indicating that the SAT is mainly a test of *g* (Frey & Detterman, 2004), although the interpretation of these findings has generated some controversy (Bridgeman, 2004; Frey & Detterman, 2005).

The theory of successful intelligence (Sternberg, 1997, 1999a) provides one basis for improving prediction and possibly for establishing greater group equity. It suggests that broadening the range of skills tested to go beyond the analytical and memory skills typically tapped by the SAT, to include practical and creative skills as well, might significantly enhance the prediction of college performance beyond current levels.

Thus, the theory does not suggest *replacing*, but rather, *augmenting* the SAT in the college admissions process. A collaborative team of investigators sought to study how successful such an augmentation could be.

1.1. The triarchic theory of successful intelligence

This study was motivated by the triarchic theory of successful intelligence (Sternberg, 1997, 1999a). Our goal was to construct-validate the theory and also to show its usefulness in a practical prediction situation. At the same time, we recognize that there are other useful theories of intelligence (see discussions in Carroll, 1993; Ceci, 1996; Cianciolo & Sternberg, 2004; Deary, 2000; Gardner, 1983; Jensen, 1998; Mackintosh, 1998; Sternberg, 1990, 2000). We are not claiming that our theory is somehow the "correct" one: no contemporary theory is likely to be final! Rather, we merely wish to

show that this theory, as operationalized, is construct valid and that it is useful in increasing predictive validity, and, at the same time, in reducing ethnic group differences in scores.

The approach we take was in many respects pioneered by Hunt (Hunt, 1980; Hunt, Frost, & Lunneborg, 1973; Hunt, Lunneborg, & Lewis, 1975), as well as by Carroll (1976), Detterman (1986), Sternberg (1977), and others (see Sternberg & Pretz, 2005, for a review of cognitive approaches to intelligence). The fundamental idea is to use modern cognitive theory to understand and measure intelligence as it pertains to school as well as other forms of success. We recognize that other approaches, such as those based on working memory (e.g., Engle, Tuholski, Laughlin, & Conway, 1999), processing speed (Neubauer & Fink, 2005), inspection time (Deary, 2000), or the combination of abilities and personality (Ackerman, Kanfer, & Goff, 1995), may ultimately prove as successful or more successful than our approach.

1.1.1. The definition of successful intelligence

The construct that forms the basis for our work is *successful intelligence*.

1. Successful intelligence is defined in terms of the ability to achieve success in life in terms of one's personal standards, within one's sociocultural context. The field of intelligence has at times tended to put the cart before the horse, defining the construct conceptually on the basis of how it is operationalized rather than vice versa. This practice has resulted in tests that stress the academic aspect of intelligence, or intelligence relevant only to the classroom, which is not surprising given the origins of modern intelligence testing in the work of Binet and Simon (1916) in designing an instrument that would distinguish children who would succeed from those who would fail in school. But the construct of intelligence needs to serve a broader purpose, accounting for the bases of success in all areas of one's life.

   We refer to the construct as successful intelligence to avoid getting into disagreements over the "true" definition of intelligence, as there is, arguably, no one true definition. Indeed, Sternberg and Detterman (1986) asked two dozen experts in the field to define intelligence, and each one gave a different definition. Our concern is with intelligence as it relates to the achievement of goals one sets for oneself within one's sociocultural context—because intelligence is a social construction. Some languages do not even have a single word for it. Our goal in this article is to propose a definition of successful intelligence, and then to operationalize that definition and test this operationalization.

   Binet and Simon (1916) originally operationalized intelligence in terms of the skills one needs for success in school. We believe this operationalization was too narrow, as indeed, did Binet, whose conceptualization was much broader than his operationalization through his test (Binet & Simon, 1916). For one thing, it would mean that intelligence is undefined for children who never go to school, and that it becomes undefined for children or adults when they leave school. For another thing, it suggests that the most important adaptation people do is to school rather than to the large majority of the years they will spend outside of it in the workforce. It is precisely against this kind of narrow and academic operationalization of intelligence that we argue. Indeed, it is this glorification of the academic experience that, in our opinion, often leads academic viewpoints to be viewed with suspicion by those outside the academy. The abilities needed to succeed in school are certainly an important part of intelligence, and they are important in the workforce (Hunt, 1995). But the abilities measured by conventional tests are not all there is, as Binet and later Wechsler (1939) both recognized in their conceptualizations of intelligence. Indeed, even intelligence, broadly defined, is only part of what is needed for success in school and in life (Sternberg, 2003). If intelligence is to be defined as what the tests test (Boring, 1923), then perhaps we at least need broader tests.

   The use of societal criteria of success (e.g., school grades, personal income) can obscure the fact that conventional operationalizations often do not capture people's personal notions of success. Some people choose to concentrate on extracurricular activities such as athletics or music and pay less attention to grades in school; others may choose occupations that are personally meaningful to them but that never will yield the income they could gain doing work that is less personally valuable. Although scientific analysis of some kinds requires nomothetic operationalizations, the definition of success for an individual is idiographic. In the theory of successful intelligence, however, the conceptualization of intelligence is always within a sociocultural context. Although the processes of intelligence may be common across such contexts, what constitutes success is not. Being a successful member of the clergy of a particular religion may be

highly rewarded in one society and viewed as a worthless pursuit in another culture.

2. One's ability to succeed requires capitalizing on one's strengths and correcting or compensating for one's weaknesses. Theories of intelligence typically specify some relatively fixed set of skills, whether one general factor and a number of specific factors (Spearman, 1904), seven multiple factors (Thurstone, 1938), eight multiple intelligences (Gardner, 1983, 1999), or 150 separate intellectual abilities (Guilford, 1982). Such nomothetic specification is useful in establishing a common set of skills to be tested. But people achieve success, even within a given occupation, in many different ways. For example, successful teachers and researchers achieve success through many different blends of skills rather than through any single formula that works for all of them. One reviewer of this manuscript suggested that our definition of successful intelligence in terms of capitalization on strengths and correcting or compensating for weaknesses is trivial—what else is there, he asked? This view, we believe, is incorrect. Positive psychology (e.g., Peterson & Seligman, 2004) emphasizes almost exclusively capitalization on strengths. Peterson and Seligman have argued that it is strengths that are important for understanding human capabilities, not weaknesses. Moreover, the traditional view in schools emphasizes correcting weaknesses–learning how to accomplish the tasks one has not mastered–rather than compensation for weaknesses—having someone else do these tasks. Indeed, compensation by seeking outside help on a test is often viewed as cheating. One is supposed to do all the work oneself. As Greenfield (1997) pointed out, only a collectivist society, such as the Maya she has studied, would view collaboration on a test as proper test-taking behavior and therefore acceptable. Indeed, intelligence is viewed differently, and manifests itself differently, in diverse cultures and other groupings (Sternberg, 2004; Sternberg, Grigorenko, & Kidd, 2005).
3. A balance of skills is needed to adapt to, shape, and select environments. Definitions of intelligence have traditionally emphasized the role of adaptation to the environment ("Intelligence and its measurement," 1921; Sternberg & Detterman, 1986). But intelligence involves not only modifying oneself to suit the environment (adaptation), but also modifying the environment to suit oneself (shaping), and sometimes, finding a new environment that is a better match to one's skills, values, or desires (selection).

   Not all people have equal opportunities to adapt to, shape, and select environments. In general, people of higher socioeconomic standing tend to have more opportunities than do people of lower socioeconomic status. The economy or political situation of the society also can be factors. Other variables that may affect such opportunities are education and especially literacy, political party, race, religion, and so forth. For example, someone with a college education typically has many more possible career options than does someone who has dropped out of high school to support a family. Thus, how and how well an individual adapts to, shapes, and selects environments must always be viewed in terms of the opportunities the individual has.
4. Success is attained through a balance of three aspects of intelligence: analytical, practical, and creative skills. Analytical skills are the skills primarily measured by traditional tests. But success in life requires one not only to analyze one's own ideas as well as the ideas of others, but also to generate ideas and persuade other people of their value. This necessity occurs in the world of work, as when a subordinate tries to convince a superior of the value of his or her plan; in the world of personal relationships, as when a child attempts to convince a parent to do what he or she wants or when one spouse tries to convince the other to do things his or her preferred way; and in the world of school, as when a student writes an essay arguing for a point of view.

1.2. Defining the three aspects of successful intelligence

According to the proposed theory of human intelligence and its development (Sternberg, 1980, 1984, 1985, 1997, 1999a), a common set of processes underlies all aspects of intelligence. These processes are hypothesized to be universal. For example, although the solutions to problems that are considered intelligent in one culture may be different from the solutions considered to be intelligent in another culture, the need to define problems and translate strategies to solve these problems exists in any culture. However, although the same processes are used for all three aspects of intelligence universally, these processes are applied to different kinds of tasks and situations depending on whether a given problem requires analytical thinking, practical thinking, creative thinking, or a combination of these kinds of thinking.

Analytical intelligence. Analytical intelligence involves skills used to analyze, evaluate, judge, or compare and contrast. It is typically used when

processing components are applied to relatively familiar kinds of problems that require abstract judgments.

Practical intelligence. Practical intelligence involves skills used to implement, apply, or put into practice ideas in real-world contexts. It involves individuals applying their abilities to the kinds of daily problems they confront on the job or at home. Practical intelligence is the application of the components of intelligence to experience to (a) adapt to, (b) shape, and (c) select environments.

Much of the work done by Sternberg et al. on practical intelligence has involved the concept of tacit knowledge. They have defined this construct as the knowledge that one is not explicitly taught and that often is not even verbalized but that one needs to work effectively in an environment (Sternberg et al., 2000; Sternberg & Hedlund, 2002; Sternberg & Wagner, 1993; Sternberg, Wagner, & Okagaki, 1993; Sternberg, Wagner, Williams, & Horvath, 1995; Wagner, 1987; Wagner & Sternberg, 1986). Sternberg et al. represent tacit knowledge in the form of production systems, or sequences of "if-then" statements that describe procedures one follows in various kinds of everyday situations (Sternberg et al., 2000). For example, if one needs to write a paper for a class (or a journal, for that matter), one can make one's way through a production system with "if...then" statements such as "If there are insufficient references on a topic, then change topics," "If the topic is too broad, then narrow it," "If the paper is too one-sided, include information on other points of view," and so on. According to this view, one essentially *constructs* one's actions by going through the production system.

Creative intelligence. Creative intelligence involves skills used to create, invent, discover, imagine, suppose, or hypothesize. Tests of creative intelligence go beyond tests of analytical intelligence in measuring performance on tasks that require individuals to deal with relatively novel situations. Sternberg has shown that assessing a range of abilities beyond that assessed by conventional tests of intelligence allows one to tap sources of individual differences measured little or not at all by these tests (Sternberg, 1985). Thus, it is important to include problems that are relatively novel in nature. These problems can call for either convergent or divergent thinking.

More details on the theory of successful intelligence and its validation can be found in Sternberg (1985, 1997, 1999a); see also Sternberg, Lautrey, and Lubart (2003).

The current study applied the theory of successful intelligence to the creation of assessments that capture analytical, practical, and creative skills. This battery was administered to more than a thousand students at a variety of institutions across the country, and was used to predict success in school as measured by GPA. The hypotheses were twofold: first, we expected that the battery of assessments based on the theory of successful intelligence would predict a substantial proportion of variance in GPA above and beyond that captured by the SAT. Second, we expected that this battery would reduce the socially defined racial and ethnic differences typically found in scores on current standardized college entrance exams such as the SAT.

2. Method

Here we outline the basic methodology used in Phase 1 of the Rainbow Project to test the hypotheses above. First, we describe the participants and institutions that participated in data collection. We then describe in detail the measures used in the study, including baseline measures and the measures we are introducing as candidates for supplementing the SAT. These measures include three multiple-choice measures from the Sternberg Triarchic Abilities Test (STAT), three practical performance tasks, and three creativity performance tasks. Finally, we conclude the Methods section with a discussion of the study design and procedure.

2.1. Participating institutions

Data were collected at 15 schools across the United States, including 8 four-year colleges, 5 community colleges, and 2 high schools.[2] Here, however, we present only the data from the colleges ($n = 13$). Most of the data were collected from mid-April 2001 through June 2001, although some institutions extended their data collection somewhat further into the summer. All institutions were supportive of our efforts to collect the data; when technical problems did occur, they tended to be with the online administration of the measures. Such technical difficulties are perhaps expected, given the fact that online data collection using these new tests of

[2] The following institutions participated in the project: Brigham Young University; Florida State University; James Madison University; California State University, San Bernardino; University of California, Santa Barbara; Southern Connecticut State University; Stevens Institute of Technology; Yale University; Mesa Community College; Coastline Community College; Irvine Valley Community College; Orange Coast Community College; Saddleback Community College; Mepham High School; and Northview High School. Students from University of California, Irvine, and from William & Mary also participated; however the *n* was less than 5 at each of these schools, so the participants were removed from subsequent analyses. In addition, 14 students failed to report institutional information, and were therefore removed from subsequent analyses for this article.

analytical, practical, and creative skills has not been done before.

2.2. Participants

Participants were recruited on a volunteer basis through fliers distributed on each campus and through psychology courses at the university and college level, and through psychology classes at the high school level. Participants either received course credit or were paid $20 for their participation.

The participants were 1013 students predominantly in their first year of college or their final year of high school. Six participants were removed from the analyses because of procedural errors, 14 students did not report institutional information, and another 3 students from 2 participating institutions were removed because the institutions did not meet the criteria for inclusion in this article (i.e., they did not have $n > 5$ students). Therefore, the total number of participants whose data were available for analyses was 990.

In this article, we include analyses[3] only for college students, except where otherwise noted.[4] Although the data from the high school students have their own utility, we analyze in detail only data from the college students because we were interested in the extent to which our new measures predict success in college, not success in high school. Thus, the final number of participants for the prediction studies presented in detail here was 777. The number of participants from each institution, their demographic characteristics, and a listing of the number of participants who completed each assessment are summarized in Tables 1 and 2. WebTable 1 is a complete by-institution-type table of means and standard deviations for all measures. See www.yale.edu/pace through May 2006, or http://pace.tufts thereafter for this and other WebTables.

2.3. Materials

2.3.1. Baseline assessments

Baseline measures of standardized test scores and high school GPA were collected to evaluate the predictive validity of current tools used for college admission criteria, and to provide a contrast for our current measures. Students' scores on standardized college entrance exams were obtained from the College Board. For most students, we accessed performance on the SAT (math and verbal sections separately, SAT-M and SAT-V), and when these scores were not available, PSAT or ACT scores were obtained. In a small number of instances where students had ACT but not SAT data, the ACT scores were transformed to match the scaling of SAT scores via the algorithm described in Dorans (1999). For the college students, high school GPA was collected from the SAT files provided by the College Board.

There is a potential concern about restriction of range in scores using the SAT when considering students from a select sample of universities. However, our sample was taken from institutions with a wide range of selectivity, from community colleges to highly selective four-year institutions. Additionally, the SD of the SAT scores (for the college sample, $SD_{SAT\text{-}V} = 118.2$, and $SD_{SAT\text{-}M} = 117.5$) was comparable with the SD of the SAT tests in the norm-group ($SD = 100$) selected to represent the broader population. If anything, a chi-squared test for differences between sample variance and population variance (Glass & Hopkins, 1996) suggests that the variance for the sample for these items is statistically larger than for the norm-group of SAT examinees (SAT-V $\chi^2(456) = 637.08$, $p < .001$; SAT-M $\chi^2(456) = 629.57$, $p < .001$). For these reasons, the concern of restriction of range of SAT scores across the whole sample is alleviated.

2.3.2. The Rainbow measures: an overview

The Rainbow measures are designed to assess analytical, creative, and practical abilities along the lines specified by the theory of successful intelligence. The instruments consisted of both multiple-choice tests (the Sternberg Triarchic Abilities Test, STAT) and performance measures of creative and practical skills. They were thus designed to sample across ability domains as well as methods of assessment.

2.3.3. The Sternberg Triarchic Abilities Test (STAT)

The STAT was developed as a means of capturing analytical, practical, and creative skills using multiple-choice questions (Sternberg & Clinkenbeard, 1995; Sternberg, Ferrari, Clinkenbeard, & Grigorenko, 1996). Level H of the test (Sternberg, 1993) was designed to measure cognitive skills among secondary school and college students, and was used in this study. The STAT briefly measures each of the triarchic skills with three types of item content: verbal, quantitative, and figural. As a result, the STAT scale is composed of nine subscales: analytical–verbal, analytical–quantitative, analytical–figural, practical–verbal, practical–

[3] Also note that psychometric scaling was done on the college sample only, unless otherwise specified.

[4] The means, medians, and standard deviations for all items for the high school students are available from the authors.

Table 1
Demographic breakdown by institution

School	Gender			Ethnicity								
	Female	Male	Total	Missing	Asian	White	Latino	Native American	Pacific Islander	Black	Other	Total
Brigham Young University	74	65	139	52	4	75	6	1	1	0	0	139
Coastline Community College	14	4	18	2	5	4	6	0	0	0	1	18
Florida State University	3	4	7	0	0	1	3	0	0	1	2	7
Irvine Valley Community College	11	7	18	2	5	8	1	0	0	0	2	18
James Madison University	31	26	57	17	0	37	1	0	0	2	0	57
Mesa Community College	77	42	119	7	3	71	14	7	1	5	11	119
Orange Coast Community College	16	6	22	2	8	2	4	1	1	0	4	22
Saddleback Community College	8	4	12	1	2	4	3	0	0	0	2	12
California State University, San Bernadino	84	27	111	19	9	37	30	0	2	9	5	111
University of California, Santa Barbara	42	17	59	9	4	37	7	0	1	1	0	59
Southern Connecticut State University	23	27	50	19	1	15	0	2	1	12	0	50
Stevens Institute of Technology	30	68	98	15	25	32	9	0	4	7	6	98
Yale University	46	21	67	12	11	25	5	0	0	10	4	67
Total	459	318	777	157	77	348	89	11	11	47	37	777

quantitative, practical–figural, creative–verbal, creative–quantitative, and creative–figural. Essay items from the STAT were not used. Each subscale included 5 items for a total of 45 items. Nine of these items (one for each of the ability × modality combinations) were new to the STAT. The particular contents of the items that compose these scales have been described elsewhere (e.g., Sternberg et al., 1996). Each multiple-choice item in the STAT had four different response options, from which the correct response could be selected. A scoring key was used for computing the STAT scores for participants who completed the tests in paper-and-pencil format. In this format, participants circled their response. The responses on the computer-administrated tests were keyed into a computer file. Ability scores were then computed by combining the responses to the subscales, using item response theory (IRT) to create three final scales representing analytical, practical, and creative skills ($STAT_{Analytical}$, $STAT_{Practical,}$ and $STAT_{Creative}$).[5] The psychometric properties of these scales are presented in the Results section.

2.3.4. Creative skills — performance tasks

In addition to the creative skill measured by the STAT, creativity was measured using open-ended measures. These performance tasks were expected to tap an important aspect of creativity that might not be measured using multiple-choice items alone, because open-ended measures require more spontaneous and free-form responses.

For each of the tasks, participants were given a choice of topic or stimuli on which to base their creative stories or cartoon captions. Although these different topics or stimuli varied in terms of their difficulty for inventing creative stories and captions, these differences are accounted for in the derivation of IRT ability estimates.

Each of the creativity performance tasks were rated on criteria that were determined a priori as indicators of creativity.[6]

Cartoons. Participants were given five cartoons, minus their captions, purchased from the archives of the *New Yorker*.[7] The participants' task was to choose three cartoons, and to provide a caption for each cartoon. Two trained judges rated all the cartoons for cleverness, humor, originality, and task appropriateness on 5-point scales. A combined creativity score was formed by summing the individual ratings on each dimension except task appropriateness, which, theoretically, is not a pure measure of creativity per se. Task appropriateness did not, and would not be expected to, correlate with the

[5] Further details on the IRT analyses are available from the authors.

[6] Further detail about the rating systems and the training of judges who rated the students' responses are available from the authors.

[7] A sample of the cartoons used in the study is available from the authors.

Table 2
Demographic data and number of college students completing each of the Rainbow measures

	College students	
	N	Percent
Gender		
Men	318	40.9
Women	459	59.1
Ethnicity		
White	348	44.8
Black	47	6.0
Latino	89	11.5
Asian	77	9.9
Pacific Islander	11	1.4
Native American	11	1.4
Other	37	4.8
Not reported	157	20.2
Completed assessments		
STAT		
Analytical	500	64.3
Practical	502	64.6
Creative	490	63.1
Creativity		
Written	441	56.8
Oral	197	25.4
Cartoons	757	97.4
Practical		
Common sense	379	48.8
College life	383	49.3
Movies	671	86.4
Year in school		
College		
First	706	90.9
Second	63	8.1
Third	6	.8
Fourth	2	.3

other ratings of creativity; however, it is a necessary prerequisite for a product to be creative with respect to a given task. In other words, a creative product is expected to be task appropriate.

Written Stories. Participants were asked to write two stories, spending about 15 min on each, choosing from the following titles: "A Fifth Chance," "2983," "Beyond the Edge," "The Octopus's Sneakers," "It's Moving Backwards," and "Not Enough Time" (Lubart & Sternberg, 1995; Sternberg & Lubart, 1995). A team of six judges was trained to rate the stories. Each judge rated the stories for originality, complexity, emotional evocativeness, and descriptiveness on 5-point scales. Because the reliability based on the total score for each story was satisfactory (see Results section), for efficiency purposes 64.7% of the stories were rated by one of the six judges.

Oral Stories. Participants were presented with five sheets of paper, each containing a set of 11 to 13 images linked by a common theme (keys, money, travel, animals playing music, and humans playing music).[8] There were no restrictions on the minimum or maximum number of images that needed to be incorporated into the stories. After choosing one of the pages, the participant was given 15 min to formulate a short story and dictate it into a cassette recorder. The process was timed by the proctor for the paper assessments and by the internal computer clock for the computer assessments. For dictation of the stories in the paper-and-pencil administration of the test, participants simply pressed the "record" button on a cassette recorder to begin dictation, and pressed "stop" when they were finished. For the computer administration, participants dictated their story into a computer microphone that translated the stories into a *.wav* file that was automatically saved onto the computer. In both cases, the actual dictation period for each story was not to be more than 5 min long. The process was then repeated with another sheet of images so that each participant dictated a total of two oral stories. Six judges were trained to rate the stories. As with the written stories, each judge rated the stories for originality, complexity, emotional evocativeness, and descriptiveness on 5-point scales. Because inter-rater reliability based on the total score for each story was satisfactory (see Results section), for efficiency purposes 48.4% of the stories were rated by only one of the six judges.

In the process of preparing this manuscript for publication, one reviewer suggested that Oral Stories may be a measure of verbal fluency rather than creativity. However, we view verbal fluency as part of creativity, and hence have no argument with this viewpoint. We agree that, at this early stage, we cannot be sure that our tests are pure measures of the constructs we seek to assess. We are hoping that the refined tests and larger sample we will use in the anticipated next phase of the Rainbow Project will help resolve such issues. Like Thurstone (1938), we think it important to separate the fluency aspect of verbal ability from its comprehension aspect (see also Carroll, 1993).

2.3.5. Practical skills — performance tasks

As outlined in Sternberg (1997), practical skills include the ability to acquire useful knowledge from experience, including "tacit knowledge" that is not explicitly taught and is often difficult to articulate, and to apply this knowledge to solving complex everyday problems. Complex everyday problems are

[8] A sample of the exact images used is available from the authors.

distinguished from academic problems in that they are practical, must be solved with incomplete information, and often do not have a single correct answer. In addition to the practical skills measured by the STAT, practical skill was assessed using three situational judgment inventories: the Everyday Situational Judgment Inventory (Movies), the Common Sense Questionnaire, and the College Life Questionnaire, each of which taps different types of tacit knowledge. The general format of tacit knowledge inventories has been described in detail elsewhere (Sternberg et al., 2000), so only the content of the inventories used in this study will be described here.[9]

Unlike the creativity performance tasks, in these practical performance tasks the participants were not given a choice of situations to rate. For each task, participants were told that there was no "right" answer, and that the options described in each situation represented variations on how different people approach different situations. That no single correct answer could be determined in our assessment situations is consistent with the kind of everyday problems that individuals with practical skills handle successfully. Even "experts" show a great deal of variability in their problem-solving strategies. The uncertainty surrounding solutions to ill-defined problem situations and the link between a particular response and resulting outcomes represents a qualitative difference between traditional cognitive testing and testing for practical skill (see Legree, 1995; Legree, Psotka, Tremble, & Bourne, 2005).

Everyday Situational Judgment Inventory (ESJI or Movies). This video-based inventory included seven brief vignettes that capture problems encountered in everyday life, such as determining what to do when one is asked to write a letter of recommendation for someone one does not know particularly well. Each situation was accompanied by six written options for how one might handle the situation. For each option, participants were asked to rate how appropriate each option was for resolving the problem on a scale from 1 (a very bad course of action) to 7 (an extremely good course of action). The ESJI took approximately 30 min to administer.

Common Sense Questionnaire (CSQ). This written inventory included 15 vignettes that capture problems encountered in general business-related situations, such as managing tedious tasks or handling a competitive work situation. Each situation was accompanied by eight written options for how one might handle the situation. Like the movie task described above, each option was rated on its quality for resolving the problem on a scale from 1 (extremely bad) to 7 (extremely good). The CSQ took approximately 30 min to administer.

College Life Questionnaire (CLQ). This written inventory included 15 vignettes that capture problems encountered in general college-related situations, such as handling trips to the bursar's office or dealing with a difficult roommate. Each situation was accompanied by several written options (with the number of options varying depending on the situation). The mean number of options for how one might handle the situation was 8. The participant indicated how characteristic and how good the option was as a means of handling the situation on a scale from 1 (e.g., not at all characteristic, not a very good choice) to 7 (e.g., extremely characteristic, a very good choice). The CLQ took approximately 30 min to administer.

2.3.6. School performance

School performance was measured using cumulative GPA as obtained from college transcripts, that is, this measure was GPA assessed at the end of the year. Clearly, GPA provides only a limited assessment of the totality of school performance. Our goal in Phase 1 of the Rainbow Project, represented here, was to see whether our measures met the minimum necessity of improving prediction of GPA.

2.3.7. Additional measures

All students at all institutions completed self-report measures of school involvement, satisfaction with school, time spent on leisure activities, competencies with computers, beliefs about the stability of cognitive skills and character, and perceptions of interpersonal competencies. These data are not presented here because preliminary analyses suggested that they did not contribute to our understanding of success in college.[10]

2.4. Design and procedure

College students filled out the assessment battery either in paper-and-pencil format (41%) or on the computer via the World Wide Web (59%).[11] Participants

[9] To avoid compromising the validity of the items in the measure, we do not present actual items used in the College Life and Common Sense measures, but instead present representative item types used in these tests. These items are available from the authors.

[10] An example item is available from the authors.

[11] The type of administration, whether paper-based or computer-based, typically depended on the institution. Because of this confound, it is difficult to determine whether there are important differences between the pencil-based versus computer-based methodologies.

were either tested individually or in small groups. During the oral stories section, participants who were tested in the group situation either wore headphones or were directed into a separate room so as not to disturb the other participants during the story dictation.

There was at least one proctor, and often two, present during the administration of the tests. Proctors read instructions to the participants for both types of administrations, and were available for questions at all times. There were two discrete sessions, conducted one after the other, for each participant. The first session included the informed-consent procedure, demographics information, the movies, the STAT items, and the cartoons, followed by a short debriefing period. The second session included obtaining consent again, followed by the rest of the demographics and "additional measures" described earlier, the Common Sense or College Life Questionnaire (depending on the condition), the Written or Oral Stories (depending on the condition), and ending with the final debriefing. The order was the same for all participants. No strict time limits were set for completing the tests, although the instructors were given rough guidelines of about 70 min per session. The time taken to complete the battery of tests ranged from 2 to 4 h.

As a result of the lengthy nature of the complete battery of assessments, participants were administered parts of the battery using an intentional incomplete overlapping design, as described in McArdle and Hamagami (1992; also McArdle, 1994). The participants were randomly assigned to the test sections they were to complete. Table 2 depicts the layout of the overlapping groups, which shows that each student completed two of the three sections of the STAT, two of the three creativity performance tasks, and two of the three practical performance tasks. The baseline (e.g., SAT-V and SAT-M) and school performance measures (e.g., GPA) were intended to be collected for all participants.

Although half the participants were to receive the oral stories, we were unable to assign the oral stories to many participants because of technical problems involving the recording equipment across different institutions. Those participants who were unable to receive the oral-stories manipulation because of these technical problems were assigned the written stories instead.

Data were also missing not only by design, but also for other reasons, mainly because of technical problems administering the tests by computer. The data that were missing for reasons other than design are listed in Table 3.

Table 3
Missing data that occurred *not* because of the intentional missing data scheme that was part of the study design

	College students	
	N	Percent
STAT		
Missing all assessments	9	1.2
Missing 1 of 2 assigned assessments	44	5.7
Practical		
Missing Movies	106	13.6
Missing both CS and CL	15	1.9
Creativity		
Missing Cartoons	20	2.6
Missing both Written and Oral	139	17.9

All missing data in the sample were managed using the full-information maximum likelihood (FIML) technique. McArdle (1994) presented the practical advantages of using FIML to estimate the parameters in structural equation model methods for handling missing data, namely, that other methods such as listwise or pairwise deletion or mean imputation result in the loss of information and potentially inaccurate computations of means and covariance data (Wothke, 2000). The particular advantage of interest here is that careful consideration of FIML during the study-design phase results in the ability to administer more assessments to a given sample where data are incomplete. Keeping the number of groups relatively constrained, unmeasured variables in a particular group (i.e., assessments not administered) can be treated as latent variables in a multigroup analysis of the entire sample including all of the measured variables (Allison, 1987; Dempster, Laird, & Rubin, 1977; McArdle, 1994; Wothke, 2000). In large-scale studies such as the one presented here, this advantage allows for the consideration of a larger number of individual differences variables.

Careful consideration should be given to the relative size of the sample groups when using FIML estimation, as large amounts of incomplete data will inflate the standard error of the estimates and reduce the power of a model (McArdle & Hamagami, 1992). That said, examples of structural equation models using FIML estimation presented in the literature feature samples with up to 80% incomplete data with generally good results for model fit and parameter estimation (e.g., McArdle & Hamagami, 1992; Wothke, 2000). The present research reflects both intentionally incomplete (i.e., by design) and unintentionally incomplete data (i.e., nonsystematic missing data), resulting in larger differences in sample groups than intended. Estimates should

thus be considered somewhat tentative. Future attempts to design our studies with intentionally incomplete data should reflect a smaller difference in sample size across groups to reach more stable estimates.

3. Results

We begin the Results section with a discussion of the descriptive statistics for the baseline assessments, namely, college GPA, and the SAT-V, SAT-M, and SAT-C. We continue with analyses of the reliability and internal factor structure of the STAT multiple-choice tests. This discussion is followed by tests of the structure of the items measuring creative abilities and the items measuring practical abilities. Following this are hierarchical multiple regressions showing the unique variance in college GPA that is accounted for by all the tests used in this study, and another multiple regression that considers a reduced number of predictors. Finally, we present an analysis of the group differences that exist for each of the measures in this study.

There is no one perfect way of analyzing these data. A major point of discussion throughout the review process of the manuscript has been with regard to the outcome variable, college GPA. As noted above, we used an incomplete design, which does not permit an execution of comparable analyses within institutions. Originally, we simply used GPAs of college students, uncorrected for the school they attended. A set of external referees used by the College Board took exception to this approach, so we corrected for level of school, using *US News and World Report* ratings of colleges as a basis for correction. But a second set of external referees used by the College Board disputed the use of this procedure. So we went back to uncorrected GPAs. Clearly, there is no perfect procedure. To satisfy the referees, we completed the analyses with both GPA and SAT standardized across the full sample of college students and GPA and SAT standardized within each college. Note that the correlation coefficient, r, for GPA_{Across} and GPA_{Within} is .91 ($p<.001$), for $SAT\text{-}V_{Across}$ and $SAT\text{-}V_{Within}$ is .69 ($p<.001$), and for $SAT\text{-}M_{Across}$ and $SAT\text{-}M_{Within}$ is .71 ($p<.001$). There were also additional points of contention, such as (a) whether SAT-V and SAT-M should be analyzed as two separate variables or as a single combined variable (SAT-C, a simple sum of SAT-V, and SAT-M) and (b) whether high school student data should or should not be a part of these analyses. All of these suggestions have been carefully considered and proper data analyses were carried out. In sum, we have analyzed the results multiple ways, and they are largely the same, regardless of method of analysis. Indeed, the general conclusions of this article hold for any of the many ways we analyzed the data over the course of our own exploratory data analyses and the various reviewers' comments. That is, no conclusions in this article change as a result of which way the data are analyzed. However, because the journal has limited space, only one set of analyses is presented here. All other analyses are available on our website (www.yale.edu/pace through May 2006, or http://pace.tufts.edu thereafter) and/or from the authors (robert.sternberg@tufts.edu).

In brief, in this manuscript, we standardize across institutions following the recommendation of the College Board (Sternberg and the Project Rainbow Collaborators, 2003); the analyses with standardization within institutions are available on the Web through WebTables. We present analyses for SAT-V and SAT-M because the test is designed, scaled, and promoted to measure different constructs (http://www.collegeboard.com/highered/ra/sat/sat.html). Although colleges often use combined SAT scores in their decision making, some liberal arts colleges put more weight on SAT-V and many schools of engineering emphasize SAT-M. Because the College Board recommends the use of separate scores rather than combined scores (Wayne Camara, personal communication, 8/6/05), we do so in our work. Yet, to satisfy the reviewers and interested readers, we share a set of analyses with SAT-C on the Web (see corresponding WebTables).

3.1. Baseline assessments

As Table 4 shows, when examining college students alone, one can see that this sample shows a slightly higher mean level of SAT than that found in colleges across the country. Using a one-sample z-test to compare the sample means with a population mean of 500 for the verbal and mathematics SAT we find statistically significant differences (for SAT-V, $z=10.19$, $p<.001$; for SAT-M, $z=14.4$, $p<.001$). The higher means in our sample may reflect that many students at these universities were recruited through their psychology courses and participated for course credit, and might capture a type of motivation that could be associated with slightly higher SAT scores overall. A more likely explanation, however, is that a relatively large proportion of the sample was enrolled in highly selective 4-year colleges. Finally, among the college students, GPA and SAT scores indicate other substantive differences, such that White and Asian students have higher GPAs and test scores than do underrepresented minority students. Group differences on these and the other measures have been found in other research (e.g., Kobrin et al., 2002) and will be discussed in detail in a later section.

Table 4
College sample descriptive statistics for GPA, SAT-V, SAT-M, and SAT-C

	GPA			SAT-V			SAT-M			SAT-C		
	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*
Institution type												
Total college sample	3.03	(0.68)	756	547.7	(118.2)	457	567.6	(117.5)	457	1115.9	(220.5)	458
2-year college only	3.05	(0.73)	183	489.6	(116.0)	48	507.9	(100.3)	48	996.2	(195.3)	47
4-year college only	3.02	(0.66)	573	554.5	(116.7)	409	574.6	(117.5)	409	1129.6	(219.3)	411
College students by gender												
Men	2.92	(0.73)	312	559.1	(113.6)	193	592.5	(110.4)	193	1151.6	(208.3)	192
Women	3.10	(0.63)	444	539.3	(121.0)	264	549.4	(119.4)	264	1090.1	(225.9)	266
College students by ethnicity												
White	3.08	(0.69)	341	576.5	(105.1)	206	589.0	(100.1)	206	1165.68	(189.6)	206
Black	2.57	(0.77)	43	498.1	(126.5)	31	506.1	(126.1)	31	1007.4	(240.0)	31
Asian	3.13	(0.64)	76	557.3	(123.7)	41	635.1	(110.0)	41	1190.8	(211.6)	40
Latino	2.97	(0.53)	86	464.5	(98.3)	53	487.2	(106.7)	53	951.7	(196.4)	53
Native American	2.40	(0.71)	10	502.5	(129.2)	4	510.0	(49.7)	4	1037.5	(146.8)	4
Pacific Islander	3.15	(0.49)	11	510.0	(62.7)	7	570.0	(89.1)	7	1080	(136.9)	7
Other	2.99	(0.60)	36	579.5	(117.4)	21	568.6	(123.0)	21	1153.81	(227.9)	21
Not specified	3.06	(0.68)	153	541.2	(127.7)	94	559.0	(129.8)	94	1101.2	(241.7)	96

Another point that should be made explicit here is that estimates of the correlations between college and high school GPA and SAT-M and SAT-V (see also WebTables for SAT-C) obtained in our study are comparable, although closer to the lower boundary, with those in the literature (Hezlett et al., 2001; Ramist, Lewis, & McCamley-Jenkins, 1994). Our correlations tend to be somewhat lower than others because we do not correct for (a) attenuation, (b) restriction of range, (c) differences in grading practices and standards across very diverse colleges and universities, or (d) reliability of the indicators in the analyses.

3.2. Sternberg Triarchic Abilities Test (STAT)

The 45 items from the STAT were analyzed together as a single test and as separate 15-item analytical, practical, and creative subtests.[12] A three-factor between-item Rasch analysis was performed on the 45-item set, representing analytical, practical, and creative constructs. In addition, to explore the verbal, numerical, and figural content of each of the 15-item subtests, the multidimensional random-coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) was applied using the ConQuest program (Wu, Adams, & Wilson, 1998).

Briefly, the Schmidt et al. analyses indicate that the IRT item reliability estimate for the 45-item STAT was good (.79). The STAT best fits a 3-factor between-item model (analytical, practical, creative) over a 1-factor model, and, when analyzed by 15-item subtest, a 3-factor model (verbal, numerical, figural) appears necessary only for the analytical subtest. Both the practical and creative subtests of the STAT appear to need only one factor to describe the dimensionality. The Cronbach alpha estimates of reliability are satisfactory but not high (.67, .56, and .72 for the analytical, practical, and creative subtests, respectively), in part because the subtests are short. The corresponding Rasch person reliability estimates for the same sample on the analytical, practical, and creative subtests were slightly lower (.59, .53, .60, respectively), which is most likely due to the presence of a ceiling effect for some particularly easy items in this test.[13] Together, these analyses support the use of

[12] The ability estimates derived from the Rasch analyses were based on the combined high school and college student sample. This approach served to increase the precision of the estimates, but did not alter in a substantive way the difference between scores of participants as the IRT estimates are sample-free (Bond & Fox, 2001; Wright & Stone, 1979). The details of these analyses, based on the combined high school and college sample (N=1013), are provided in the technical report prepared independently by the Jefferson Psychometric Lab (Schmidt, Bowles, Kline, & Deboeck, 2002).

[13] Within the many-facets Rasch model, responses are modeled at the item level. Thus, each item and person has a corresponding standard error, which allows for a more accurate computation of reliability than a simple Cronbach's alpha, which is determined based on the error of a hypothetical "average" test taker. Ceiling effects are likely to cause item response patterns that are too consistent, resulting in low infit scores and low person separation. Cronbach's alpha is less sensitive to these effects. These data, available from the authors, show the distribution of scores for each of the STAT subscales, as well as their correlations with college GPA.

separate subtest scores. In subsequent analyses, the IRT ability estimates for the analytical, practical, and creative subtests based on both high school and college students are used in preference to the raw scores.

3.3. Creative abilities — performance tests

3.3.1. Cartoons[14]

As described above, the cartoon task was scored along multiple dimensions, including *cleverness*, *humor,* and *originality*, leaving out *task appropriateness* because the responses were all largely appropriate to the task. Using facets analysis (an extension of the Rasch model), we derived a single ability estimate related to the cartoons for each participant (Linacre, 1989). The four measurement facets used were the same as for the Written and Oral Stories: person ability, item dimension differences (i.e., for cartoons the dimensions of this facet used were: cleverness, humor, and originality), rater severity, and story difficulty. Using the many-facets Rasch model for analysis has two distinct advantages in this context. First, the ratings from multiple judges may be accurately combined into a single score. Second, each of the facets under examination may have its elements compared on a common scale (i.e., the logit scale). Thus, we can get an empirical estimate of which items were most difficult, which raters were most severe, which item dimensions were most difficult, and which students had the highest ability. Because the estimates for each facet are on a common scale, person ability estimates can then be accurately adjusted for differences in the severity of the judges scoring each person's items, the difficulty level of the items that were selected, and the difficulty level for each of the creativity dimensions. This yields a single overall ability estimate for each student that has been adjusted for each of the facets in the model.

Table 5 reports the zero-order correlations between these ability-based scores (not adjusted for selectivity) and shows evidence that our judges were able to differentiate task appropriateness from other measures thought to capture creativity. The IRT reliability for the composite person ability-based estimates was very good (CHO = .86). The results also indicate slight differences in the level of severity between the raters; however, all raters fit the model very well, such that any differences between raters could be reliably modeled (reliability = .99). Finally, the results indicate that the range in difficulty from one cartoon to the next was small (− .16 to .18). Therefore, fit statistics indicate that each of the items fit the model well, and that the variance in difficulty could be reliably modeled (reliability = .96; see Schmidt et al., 2002, for an independent report on item analyses).

Table 5
Intercorrelations between creativity components (Rasch estimates) of the Cartoons Task

	C	H	O	TA	CHO
Cleverness (C)	1.00				
Humor (H)	0.82	1.00			
Originality (O)	0.76	0.72	1.00		
Task appropriateness (TA)	0.39	0.41	0.23	1.00	
Composite (CHO)	0.93	0.92	0.90	0.37	1.00

$N = 757$.

3.3.2. Written and Oral Stories

The raw scores assigned by the raters were analyzed by the many-facets Rasch model (FACETS; Linacre, 1989, 1994) using the FACETS computer program (Linacre, 1998).[15] Four measurement facets were used: person ability, item dimension differences, rater severity, and story difficulty. Student ability estimates were derived for the *complexity*, *emotionality*, *descriptiveness*, and *originality* dimensions on which the responses were rated. Table 6 reports the zero-order correlations between these components for each task. The Rasch reliability indices for the composite person ability estimates for the Written and Oral Stories were very good (.79 and .80, respectively). The judges for both the Written and Oral Stories varied greatly in terms of their severity of ratings for the stories. For the Written Stories, the judges also ranged in their fit to the model, although the reliability was still sound (rater reliability = .94).

For the Oral Stories, all the judges fit the model very well, so their differences could be reliably modeled (rater reliability = .97). Finally, the results indicate that differences between the choice of story titles for the Written Stories and images sheets for the Oral Stories were modest (− .15 to .14 for the Written

[14] The ability estimates derived from the Rasch analyses were based on the combined high school and college student sample. This approach served to increase the precision of the estimates, but did not alter in a substantive way the difference between scores of participants as the IRT estimates are sample-free (Bond & Fox, 2001; Wright & Stone, 1979).

[15] The ability estimates derived from the Rasch analyses were based on the combined high school and college student sample. This approach served to increase the precision of the estimates, but did not alter in a substantive way the difference between scores of participants as the IRT estimates are sample-free (Bond & Fox, 2001; Wright & Stone, 1979).

Table 6
Intercorrelations between components of Written (A) and Oral (B) Stories (Rasch estimates)

A. Written Stories

	CO	EM	DE	OR	WS
Complexity (CO)	1.00				
Emotionality (EM)	0.77	1.00			
Descriptiveness (DE)	0.63	0.56	1.00		
Originality (OR)	0.35	0.33	0.29	1.00	
Composite Written Stories (WS)	0.82	0.78	0.76	0.58	1.00

N=441

B. Oral Stories

	CO	EM	DE	OR	OS
Complexity (CO)	1.00				
Emotionality (EM)	0.68	1.00			
Descriptiveness (DE)	0.61	0.48	1.00		
Originality (OR)	0.46	0.37	0.29	1.00	
Composite Oral Stories (OS)	0.82	0.74	0.75	0.65	1.00

N=197

Stories, and − .15 to .10 for the Oral Stories), such that differences could be reliably modeled (reliability for Written Story titles=.91, for Oral Story images=.81). Further, independent item analyses are reported in detail by Schmidt et al. (2002). One conclusion from the item analyses is that the originality component needs some refinement, at least in terms of scoring; however, removing the originality scores did not substantially change the reliability of the measures. Therefore, the ability estimates for the Written and Oral Stories were based on the four rated dimensions, each of which was used in subsequent analyses.

3.3.3. Latent factor structure of performance measures of creativity

The experimental design does not allow direct comparison of the relationship between the creativity measures because participants received either the Oral Stories or the Written Stories, but not both; however, all participants received the Cartoons task (see Method section). Therefore, the covariance matrix for these measures was estimated using the full-information maximum likelihood (FIML; Allison, 1987; Dempster et al., 1977; McArdle, 1994) method as implemented in Mplus version 3.13 (Muthen & Muthen, 2002). The estimation algorithm is assisted by additional variables that have overlapping samples of respondents, and so the standardized college GPA, high school GPA, the SAT-V and SAT-M, as well as the $STAT_{Creative}$ measures were included in the analyses. The estimated correlation matrix for these variables is provided in Table 7. Note that Oral Stories and $STAT_{Creative}$ correlate almost as highly or higher with college GPA as the SAT-V and SAT-M (Table 7) or SAT-C (WebTable 7) do with high school GPA.

The Rasch analyses suggest that the separate performance measures of creativity have appropriate internal psychometric properties. However, the intercorrelations between pairs of the creativity tasks are themselves quite small, suggesting that the possibility of identifying a single common latent factor uniting these variables (i.e., variables 2, 3, 4, and 5 in Table 7 for SAT-V and SAT-M and WebTable 7 for SAT-C) is low. As has been stated elsewhere, creativity is, at least in part, domain specific (Sternberg, Grigorenko, & Singer, 2004; Sternberg & Lubart, 1995). It becomes domain general only when measured solely in the most trivial ways (such as through very simple fluency measures).

The model summarized in Fig. 1 explores the incremental prediction of such a latent creativity factor and reports the fit statistics, standardized path coefficients and their standard errors, and the squared multiple correlation between college GPA and all variables in the model. The overall fit is good ($\chi^2(9)=16.74$, $p=.053$, CFI=0.970, RMSEA=.033, 90% CI=.000–.058), based on standard criteria of CFI indices larger than .95 and RMSEA indices equal to or lower than .05. The creative component of the STAT contributed significantly to the incremental prediction of college GPA; however, the composite performance measure of creativity did not. There are substantial correlations between the SAT-V and the latent creativity factor,

Table 7
Estimated correlations between creative abilities, SAT, and high school and college GPA

	1	2	3	4	5	6	7
1. College GPA[a]	1.00						
2. Oral Stories	.28	1.00					
3. Written Stories	.13	.07	1.00				
4. Cartoon	.08	.16	.23	1.00			
5. $STAT_{Creative}$	.34	.08	.30	.28	1.00		
6. SAT-V[a]	.27	.22	.37	.38	.54	1.00	
7. SAT-M[a]	.29	.19	.29	.27	.59	.75	1.00
8. High school GPA[a]	.37	.05	.20	.20	.46	.50	.57

Nominal N=777; the nominal n of 777 represents the n of all students taking any portion of any test. There are no students who took every portion of every test. When the FIML procedure is used, the correlation matrix is estimated based on all of the information available from all tests for all participants. Rather than using pairwise or listwise deletion techniques to compute correlations, the FIML technique computes pairwise correlation values, and then adjusts the value of the correlation based on the information from other variables in the model. Further information on the FIML procedure may be found in McArdle (1994) and McArdle and Hamagami (1992).

[a] z-score transformation applied.

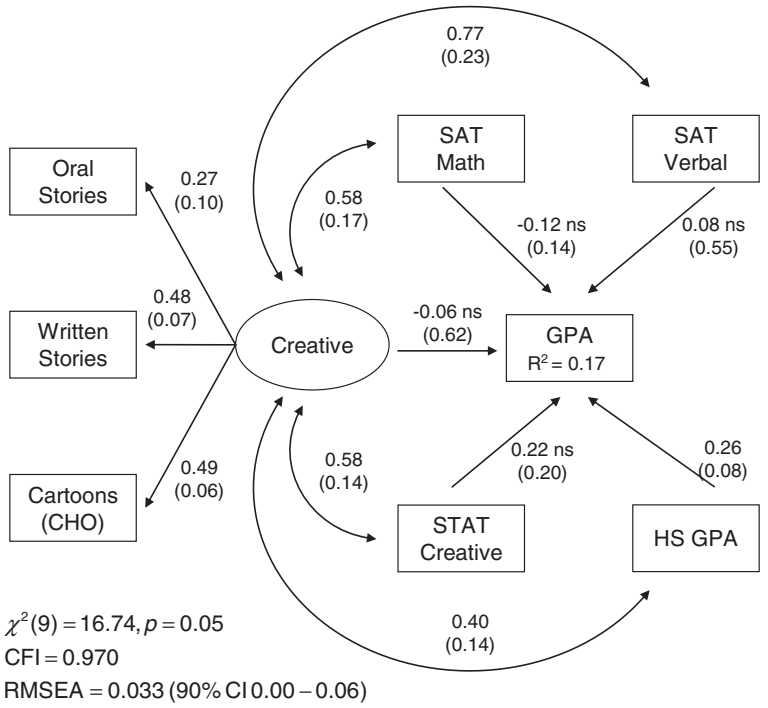

Fig. 1. Prediction of College GPA (GPA) by Creative Abilities, SAT-V and SAT-M, and High School GPA (HSGPA).

which suggests that, for this sample, the performance measures of creative ability and the verbal test of the SAT are tapping common content to an important extent. The zero-order correlations between the SAT-V and the Oral, Written, and Cartoon measures are .22, .37, and .38, respectively, whereas the correlation between SAT-V and $STAT_{Creative}$ is .54 (see Table 7). WebFig. 1 presents this model with SAT-C.

3.4. Practical abilities — performance tests[16]

For all three tacit knowledge measures, the scores assigned to each participant were derived by calculating the Mahalanobis distance (D^2) of the participant's ratings for each possible solution strategy from the mean ratings of the sample to which the participant belonged. A brief description of the calculation of D^2 follows, using the Everyday Situational Judgment (Movies) Inventory as an example to illustrate how the calculation was done (see also Rencher, 1995). The same procedure was used on the Common Sense Questionnaire and the College Life Questionnaire, and the following example could be applied to those tests as well.

For each of the six possible solution strategies accompanying each of the seven vignettes in the Everyday Situational Judgment Inventory (Movies), the sample's mean rating (excluding the rating of the participant of interest) was subtracted from the participant's rating. These computations resulted in a vector of six simple difference scores for each participant, for each of the seven vignettes, and thus $7 \times N$ vectors in all. Then, the vectors of difference scores were each multiplied by the inverse of the variance–covariance matrix of the six possible response strategies from which the difference scores were created. The resulting 6×1 vector was then multiplied by the transpose of the original difference-score vector, resulting in a scalar, called the Mahalanobis distance, or D^2. These computations, then, resulted in seven D^2 values per individual, one per vignette, and thus $7 \times N$ in all. The D^2 values were then averaged, and their square root was taken to return the value to its original metric. The individual's total score for the Everyday Situational Judgment

[16] The scaling for all of the practical tests was done using the college student sample only. The reason for this approach is because the practical measures were scored using a group-based scoring approach rather than an item response theory approach (as was used in scoring the creative tasks).

Inventory (Movies) was determined by averaging the resulting vignette-level values.[17]

Situational judgment inventories, used in personnel research for decades, traditionally feature a set of response options from which the examinee is asked to select either the best response or the best and worst response (Legree, 1995; Legree et al., 2005; Motowidlo, Hanson, & Crafts, 1997). The use of a Likert-type scale for rating the quality of response options and a distance-score methodology for determining relative performance levels has been shown to improve the reliability and construct validity of situational judgment inventories for assessing interpersonal skills (Legree, 1995; Legree et al., 2005). Although the use of Mahalanobis distance scores to indicate practical abilities is a novel application of this statistic, it represents an extension of earlier work and is logically consistent with the use of D^2 to detect outliers in multivariate distributions. Yet, to ensure the consistency of our results, we also computed distances at the option level. For example, the movies had 7 scenarios with 6 options each, resulting in 42 responses at the option level. Correspondingly, whereas the Mahalanobis distance scores utilize only summative information from the 7 scenarios, absolute deviation values obtained at the option level utilize all available information from all responses. Unsurprisingly, the reliability estimates at the option level tend to be higher than at the scenarios/vignettes level. Below we present reliability estimates for both types of scoring and the correlations between the scoring approaches. Because the correlations are very substantial, we resort to the use of the more conceptually appropriate, from our point of view, method of scoring with the Mahalanobis distances.

As noted above, scores on the practical ability performance measures were determined in reference to the average, or consensual, responses of the sample. Important concerns arise when consensual scoring techniques become imbalanced with regard to race, ethnicity, or sex, as such imbalances might be biased against minority group members; other problems arise with regard to defending the basis of any particular individual's score against the average responses of the sample. However, using an "expert group" as a reference instead of the average responses of the sample might lead to similar problems, for example, with determining the demographic characteristics of those individuals comprising such an "expert group." Legree (1995) demonstrated that the ratings of experts and nonexperts on a situational judgment inventory were highly correlated (r=.72 and .95), indicating that a fairly knowledgeable nonexpert consensus was as sensitive to relative differences in solution quality as were the experts. Mayer, Salovey, Caruso, and Sitarenios (2003) have shown that an expert panel shows more *within-group* consistency than a general sample in selecting the "correct" answer on emotional intelligence items; however, there appears to be a great deal of *between-group* agreement in terms of these items, suggesting that both expert panels and general samples tend to agree on the overall correct answers to emotional intelligence items.

3.4.1. Everyday Situational Judgment Inventory (Movies)

Of the 777 college students included in the analysis for this study, 670 produced complete data, and two students produced usable but incomplete data. Missing data were a result of technological difficulties in either showing the films or collecting data via computer. Two exceptions were participants who were removed from analyses for apparent malfeasance (e.g., rating all 42 response options with a "1"). *All* of the items on the test require procedural rather than factual/declarative knowledge to be answered correctly. That is, there are no problems that can be answered on the basis of declarative knowledge alone because all require problem solving, even if declarative knowledge is used in such problem solving.

3.4.1.1. Measurement properties. The internal-consistency reliability of a scale composed of the seven distance scores was determined using Cronbach's alpha. This reliability was .76 for the Mahalanobis distance (D^2) and .80 for absolute deviation values (for the two scoring approaches, r=.95, p<.001), which is comparable with that of many conventional ability tests containing more items.

3.4.1.2. Underlying structure. Consistent with Wagner (1987), the fit of a single-factor model to the data was tested via confirmatory factor analysis (CFA). The fit of this model was very good ($\chi^2(14)=21.65$, p=.09; CFI=.99; RMSEA=.03, 90% CI=.00–.05), with loadings of the vignettes on the latent factor ranging between .50 and .60. The variance accounted

[17] Although there is a conceptual difference between using the Euclidian distance measure (d^2) and the Mahalanobis distance measure (D^2), the results were run using both approaches. The correlation between the two sets of distance measures was greater than 0.97 for the entire sample.

for in the vignettes by the latent factor (R^2) ranged between .25 and .36, however, indicating that the vignettes could be improved in their measurement of practical abilities as represented in the acquisition and use of general, everyday tacit knowledge. Although the shared variance among the vignettes could be accounted for by a single factor, much unique variance remained. To some degree, unique variance should be expected, as each vignette features a different problem situation. Furthermore, the commonalities (amount of common variance) for these vignettes are comparable with, if not higher than, those reported for measures of cognitive abilities as traditionally defined (e.g., Cattell's Culture Fair Test of *g*, Engle et al., 1999; Arithmetic Reasoning, Kyllonen & Christal, 1990; Raven's Progressive Matrices, Rogers, Hertzog, & Fisk, 2000) or for working memory (Alphabet Recoding, Kyllonen & Christal, 1990; Operation Span, Reading Span, and Computation Span, Engle et al., 1999) in CFAs where they were specified to load on a single, construct-relevant higher-order factor (e.g., *gf* or general working memory). The results of this analysis justified the formation of a single composite for further analyses, representing practical abilities as reflected in the acquisition of general, everyday tacit knowledge. This composite was formed by taking the unit-weighted average of the Mahalanobis distances across all 7 vignettes.

3.4.2. Common Sense Questionnaire

Roughly half of the 777 college students included in these analyses ($n=377$) produced complete data for the College Student Questionnaire as a result of the intentional incomplete overlapping-group design described earlier (McArdle, 1994). Three participants were removed from analyses for apparent malfeasance.

3.4.2.1. Measurement properties. The internal-consistency reliability of a scale composed of the 15 vignettes was determined using Cronbach's alpha. This reliability was .91 for D^2 and .95 for the absolute deviation values ($r=.93$, $p<.001$), which is comparable with that of many conventional ability tests.

3.4.2.2. Underlying structure. As with the data from the video-based vignettes, a CFA was used to test the fit of a single-factor model to the data. The fit of this model was good ($\chi^2(90)=217.91$, $p=.00$; CFI=.94; RMSEA=.06, 90% CI=.05–.07), with loadings of the vignettes on the latent factor ranging between .58 and .70. The commonalities for the vignettes ranged between .34 and .49, indicating that the vignettes are reasonable measures of practical abilities as reflected in the acquisition of general, business-related tacit knowledge. Once again, the vignettes appear to be comparable in structure with conventional measures of cognitive abilities or working memory. The results of this analysis justified the formation of a single composite for further analyses. This composite was formed by taking the unit-weighted average of the Mahalanobis distances across all 15 vignettes.

3.4.3. College Life Questionnaire

Roughly half of the 777 college students included in these analyses ($n=385$) were administered the College Life Questionnaire as a result of the intentional incomplete overlapping-group design. Ten participants were removed from analyses for apparent malfeasance or failure to follow directions.

3.4.3.1. Measurement properties. The internal-consistency reliability of a scale composed of the 15 vignettes was determined using Cronbach's alpha. This reliability was .89 at the vignette level and .95 at the option level ($r=.93$, $p<.001$), again comparable with that of many conventional ability tests.

3.4.3.2. Underlying structure. As with the data from the video-based vignettes and the Common Sense Questionnaire, a CFA was used to test the fit of a single-factor model to the data. The fit of this model was marginal ($\chi^2(90)=244.42$, $p=.00$; CFI=.92; RMSEA=.07, 90% CI=.06–.08). The loadings of the vignettes on the latent factor ranged between .53 and .70, with the exception of one vignette that had a loading of .38. The majority of the commonalities for the vignettes ranged between .28 and .49, indicating that the vignettes are reasonable measures of the underlying construct of practical abilities as reflected in the acquisition and use of general, college-related tacit knowledge. The vignette with the relatively low loading, whose commonality was .14, appears to be an exception.

Examination of the content of this vignette reveals that participants were asked to indicate their preference for particular school-related activities, rather than rate the quality of the activities as a strategy for achieving a particular goal. As tacit knowledge is applied toward achieving a particular goal (adapting to, shaping, or selecting the environment), this vignette did not capture a key aspect of using practical abilities. The relatively poor measurement of the construct by this vignette suggested that it should be removed from further analyses. The same CFA when fit to the data with this

variable excluded had acceptable fit ($\chi^2(77)=205.814$ $p=.00$; CFI=.92; RMSEA=.07, 90% CI=.06–.08). The results of this analysis justified the formation of a single composite composed of 14 items for further analyses, representing practical abilities as reflected in the acquisition of general, college-related tacit knowledge captured across vignettes. This composite was formed by taking the unit-weighted average of the Mahalanobis distances across all 14 vignettes.

In summary, all three indicators of practical ability have adequate internal consistencies and concur with the anticipated theoretical structure. In fact, the reported reliabilities compare favorably with those for many situational judgment tasks (SJTs). For example, in a meta-analysis of SJTs, McDaniel et al. (McDaniel, Morgeson, Finnegan, Campion, & Braveman, 2001) reported a median reliability of .795, with a range of values from .43 to .94, with half the values below .80 and one-third below .70. As per Nunnally's (1978) recommendation of viewing a reliability value of .80 as a minimum level for applied projects, and .70 for basic research, our indicators of practical ability meet the required standards.

3.4.4. Practical abilities, SAT, and GPA

The intercorrelations between the vignette-based practical ability measures, the $STAT_{Practical}$ composite, SAT-V, SAT-M, and GPA are shown in Table 8 (see WebTable 8 for SAT-C). These intercorrelations were estimated for nearly the entire sample of college students ($N=777$) in Mplus using FIML estimation.

These intercorrelations indicate that practical abilities, as reflected in the acquisition of tacit knowledge of differing content, show some relation to GPA. This finding is particularly true for the Common Sense Questionnaire, which assesses general business tacit knowledge. Surprisingly, the College Life Questionnaire and the Everyday Situational Judgment Inventories (Movies), which depicted problem situations typically experienced during undergraduate education, showed relatively smaller relations to college GPA.

Table 8
Estimated correlations between practical abilities, SAT-V, SAT-M, and high school and college GPA

	1	2	3	4	5	6	7
1. College GPA[a]	1.00						
2. Everyday Situational Judgment	.14	1.00					
3. College Life Questionnaire	.15	.59	1.00				
4. Common Sense Questionnaire	.27	.54	.31	1.00			
5. $STAT_{Practical}$	.25	.28	.28	.36	1.00		
6. SAT-V[a]	.28	.28	.24	.27	.54	1.00	
7. SAT-M[a]	.29	.26	.26	.30	.57	.75	1.00
7. High school GPA[a]	.37	.17	.23	.26	.45	.50	.57

Nominal $N=777$; FIML used to estimate statistics.
[a] z-score transformation applied.

Next, a full structural equation model was fit to the data to examine the simultaneous relations between these measures and college GPA. As with the intercorrelations presented previously, the estimates presented in this model were derived using data from nearly the entire sample ($N=777$) and FIML estimation. Fig. 2 shows the model and the corresponding fit indices, standardized path coefficients, and correlations. The fit of this model is good, as indicated by a nonsignificant χ^2, a CFI of .99, and an RMSEA of .02, whose 90% confidence interval contains .00. As shown in Fig. 2, the three tacit knowledge measures each load highly on a single general practical abilities factor, as expected. Because of a compromise to model fit that occurs when the $STAT_{Practical}$ items are included with the performance items, it was not specified to load on the practical latent variable. These problems may occur because the STAT scale uses a different methodology compared with the other practical performance measures. Nevertheless, there is a significant correlation between the $STAT_{Practical}$ and the latent variable comprising the practical performance items, suggesting that each is tapping a similar construct.

Importantly, the general practical abilities factor shows a significant path coefficient to college GPA, the only significant path coefficient to college GPA other than high school GPA. SAT-M and SAT-V, when analyzed simultaneously with general practical abilities, do not significantly account for variance in GPA. Both SAT-M and SAT-V show a significant relation to the practical latent variable. In the case of SAT-V, this relation may occur because general practical abilities in our tests are indicated by measures that require verbal processing to complete. The relation between SAT-M and the general practical ability factor was not expected, but could perhaps be explained by a reasoning or problem-solving component that may be common to both types of measures. Nevertheless, overall, the fit of this model and the significant path from general practical ability to college GPA show some promise in using vignette-based practical ability measures to supplement SAT scores when considering candidates for college admission. WebFig. 2 presents this model with SAT-C.

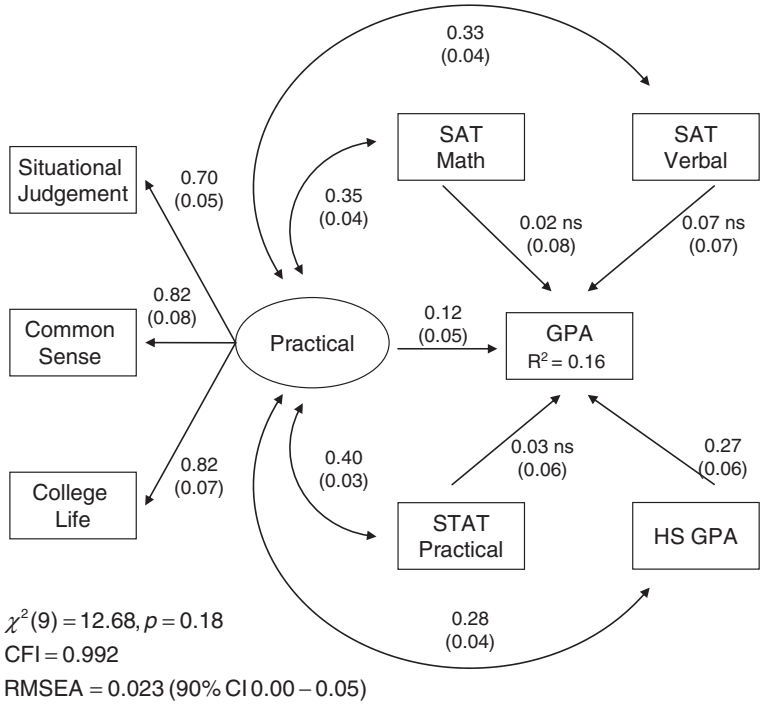

Fig. 2. Prediction of College GPA (GPA) by Practical Abilities, SAT-V and SAT-M, and High School GPA (HSGPA).

3.5. Factor structure of the triarchic measures

An exploratory factor analysis was conducted to investigate the factor structure underlying the triarchic measures. The results of these analyses are reported in Table 9. Specifically, because the three factors are not theorized to be completely orthogonal, a promax rotation was performed. Three factors were extracted with eigenvalues greater than 1 and these accounted for 62.8% of the variation among the measures. Table 9 presents the pattern matrix, as well as the intercorrelations among the three factors.

Table 9
Exploratory factor analysis of triarchic measures

	Estimated correlations[a]								F1	F2	F3
	1	2	3	4	5	6	7	8			
1. Oral Stories	1.00								**.57**	−.06	−.06
2. Written Stories	.07	1.00							**.79**	.01	−.02
3. Cartoons	.14	.24	1.00						**.20**	**.28**	−.08
4. $STAT_{Creative}$	.11	.27	.29	1.00					.00	**.73**	.09
5. $STAT_{Analytical}$	.14	.24	.21	.58	1.00				−.06	**.80**	−.04
6. $STAT_{Practical}$	.14	.31	.29	.61	.63	1.00			.03	**.81**	−.02
7. Movies	.02	.22	.14	.29	.17	.26	1.00		.12	.05	**.52**
8. College Life	.01	.13	.12	.38	.24	.30	.59	1.00	−.13	.01	**1.00**
9. Common sense	.03	.30	.05	.38	.38	.33	.55	.33	.12	−.01	**.92**

		Factor intercorrelations		
		F1	F2	F3
	F1	1.00		
	F2	.45	1.00	
	F3	.30	.40	1.00

62.8% of variation explained; Nominal $N = 776$. Bolded numbers indicate salient loadings on factor
[a] correlations estimated using FIML.

The results suggest that, consistent with the analyses reported above, evidence for a unidimensional latent creativity factor is unclear, although the four creativity indicators load on this factor with coefficients ranging from .20 to .79. The practical ability measures clearly define a latent factor, again consistent with the analyses reported above. That the STAT variables define a latent factor is expected to the extent that methodology (multiple choice) and to some extent content (numerical, figural, verbal) is common across the analytical, practical, and creative items. It would seem that, in this sample, the common methodological factor might overwhelm the unique creative, practical, and analytical contribution offered by the different STAT subtests.

3.6. Model comparisons

When using structural equation models to evaluate predictive validity, it is important to compare competing explanatory models. Thus, Table 10 shows a comparison of five potential explanatory models.

Model 1 is the standard *g*-based model, which specifies that each of the manifest variables in the model loads on a common general ability factor. This *g*-factor is then used to predict freshman GPA. According to the results listed in Table 10, the *g*-based model was not a good fit to the data, as indicated by a statistically significant χ^2, a CFI of .83, and an RMSEA of .080, whose 90% confidence interval does not contain .05, suggesting the data do not support this model.

Model 2 is indicated as a strict triarchic model, in which all manifest variables were forced to load on one of three latent variables (i.e., analytical, creative, or practical). Although this model was a slight improvement over the *g*-based model, the strict triarchic model was not a good fit to the data as indicated by a statistically significant χ^2, a CFI of .86, and an RMSEA of .076, whose 90% confidence interval does not contain .05.

Model 3 specifies a "method model" in which the latent variables simply represent method variance. Thus, those variables that were measured using some form of multiple-choice assessment were specified to load on the multiple-choice latent variable, and those manifest indicators that were measured in a performance-oriented way were specified to load on a performance latent. High school GPA did not fit well into either of the latent method indicators, and therefore was also included in the model as a manifest variable independently of the two latents. The fit for this model was better than the first two models, but was again not very good by traditional SEM criteria. The χ^2 test is statistically significant, the CFI is .89, and the RMSEA is .065, with a 90% confidence interval that does not contain .05.

Model 4 specifies three latent variables. The first is composed of the performance measures of creative ability. The second is composed of the performance measures of practical abilities, and the third latent variable is the STAT, viewed in this model as largely a *g*-based (i.e., analytical) test. This model showed a substantial improvement in fit over all previous models, and indeed showed good fit by standard SEM criteria. Model 4 had a nonsignificant chi-squared statistic, a RMSEA of .030, and a CFI of 0.98.

Finally, Model 5 shows a modified version of Model 4. The key difference between these two models is that in Model 5, each of the three components of the STAT are kept as manifest indicators that do not load on any latent factors. In this way, the model reflects that the STAT does not measure *only g*. Model 5 also showed very good fit to the data, although it did not show substantial improvement over and above Model 4.

In summary, Models 4 and 5 showed much better fit to the data than did any of the first three models. Therefore, in the next section, the regression equations are presented for the subcomponents presented in Model 5.

3.7. Complete hierarchical regressions

Regressions are described below. None are corrected for restriction of range, attenuation, or shrinkage; hence these figures are not directly comparable with those of other investigators who have done such corrections.

Many investigators correct correlation coefficients for restriction of range. We do not, for several reasons: (a) we think the assumptions underlying corrections are somewhat dubious, and (b) we do not know what the mean and standard deviation for creative and practical tests for this population would be, as no norming studies exist for these measures.

Many investigators also correct correlation coefficients for attenuation. We do not, although not because we have any great objection in principle to doing so. But we believe that such corrections once again entail dubious assumptions, at times even leading to correlations greater than 1. Moreover, the lower the reliability of the test and the greater the correction, the less likely it seems to be to represent reality. There is, of course, value in corrections, and some investigators prefer corrected correlations. But we believe the greatest value is in reporting what the data were, not what they might

Table 10
Structural equation model comparisons

	Model	Chi-squared	*df*	RMSEA	90% CI	CFI	R^2
Model 1	*g* model	377.60	63	.080	.072–.088	.827	.136
Model 2	Strict triarchic [a]	319.31	58	.076	.068–.084	.856	.172
Model 3	Method [b] + hsgpa model	251.60	59	.065	.057–.073	.894	.164
Model 4	Perf_C [c] + Perf_P [d] + STAT + zhsgpa + zsat_m + zsat_v	78.34	46	.030	.018–.041	.975	.177
Model 5	Perf_C [a] + Perf_P [a] + STATcre + STATanl + STATpra + zhsgpa + zsat_m + zsat_v	53.94	34	.027	.012–.041	.970	.178

[a] Creative = oral, written, cartoon, $STAT_{Creative}$; Practical = movies, college, common, $STAT_{Practical}$; Analytic = sat_m, sat_v, hsgpa, $STAT_{Analytic}$.
[b] Multiple-choice = STAT analytic, $STAT_{Creative}$, $STAT_{Practical}$, SAT-V, SAT-M; Performance = oral, written, cartoons, movies, college, common.
[c] Perf_C = cartoons, oral, written.
[d] Perf_P = movies, college, common.

have been had certain (sometimes dubious) assumptions been met.

3.7.1. Predicting college GPA[18]

To test the incremental validity provided by triarchic measures above and beyond the SAT in predicting GPA, a series of hierarchical regressions was conducted that included the items analyzed above in the creative and practical abilities. To complete the third dimension of the triarchic model, we also included the $STAT_{Analytic}$ measure. The estimated correlation matrix on which these analyses are based is provided in Table 11.1.[19] (see WebTable 11 for correlation coefficients computed using the SAT-C indicator). Table 11.2 provides the calculated correlation matrix. The hierarchical regressions that include all three dimensions of the triarchic model are shown in Tables 12 and 13. Note that the creativity measures in these hierarchical regressions are separated from their latent variable because, as noted earlier, these items did not include enough common variance.

As shown in Table 12 (1 and 2), SAT-V, SAT-M, and high school GPA were included in the first step of the regression because these are the standard measures used today to predict college performance. Here SAT and GPA indicators were standardized across institutions. (See WebTables 12 A–N for parallel analyses with indicators standardized within and across institutions in difference combinations and with SAT-C.)

Only high school GPA contributed uniquely to R^2. In Step 2 we added the analytical subtest of the STAT, because this test is closest conceptually to the SAT tests. The inclusion of the analytical subtest of the STAT did not contribute to the explained variance, and in fact, suggests the presence of a trivial suppressor effect, indicating that it had nothing substantive to contribute and may have been capitalizing on chance or minor variations in the data. In Step 3, the measures of

[18] One problem when using college GPA from students across different colleges is that a high GPA from a less selective institution is equated to a high GPA from a highly selective institution. One could make the argument that the skills needed to achieve a high GPA at a selective college are greater than the skills needed to achieve a high GPA at a less selective college. There are a number of ways one could account for this problem of equated GPAs. (1) One could assign a weight to GPA based on the selectivity of the students' institution, such that more selective institutions are given a weight that increases the GPA relative to less selective institutions. However, this procedure assumes that the variables used to predict GPA are measured independently of the weight, namely selectivity of the school. Because SAT is used to determine the selectivity of the school to which a student matriculates, and therefore results in a violation of independence of independent and dependent variables, we could not run this procedure because it would artificially inflate the relationship between SAT and weighted GPA. Adjusting for the SAT/Selectivity relationship by partialling out selectivity from the SAT would artificially deflate the relationship between SAT and weighted GPA. (2) A second procedure would be to standardize all scores, including the dependent variable and all independent variables, within levels of selectivity of the institution, or even within each school, and then run these scores together in all analyses. This standardization procedure effectively equates students at highly selective institutions with students from less selective institutions, and produces results that would be essentially a rough summary of the analyses done within each level of selectivity or within each school. One problem with this procedure is that it loses the elegance of involving schools in a large range of selectivity (e.g., University of California at Santa Barbara versus Mesa Community College) if all students become equated by standardization. Nevertheless, when this procedure is run, the pattern of results is essentially the same as an analysis that does not use a standardization adjustment to the data; in fact, the only substantive change is that, across the board, all coefficients become attenuated (including correlations, beta coefficients, R^2, et cetera). Consequently, we have chosen to report the results based on scores that are unadjusted for institutional selectivity.

[19] The estimation of correlations in FIML is partially dependent on the variables included in the model. This results in minor differences in the estimated values as will be noted for instance in the comparison of Table 9 with Table 10. For comparison with the raw correlations of the measures without FIML computation (i.e., with incomplete samples), contact the authors.

Table 11.1
Intercorrelations between Rainbow measures, GPA, and SAT

	Mean	1	2	3	4	5	6	7	8	9	10	11	12
1. College GPA[a]	.00	1.00											
2. High school GPA[a]	−.04	.36	1.00										
3. SAT-M[a]	−.12	.28	.57	1.00									
4. SAT-V[a]	−.11	.26	.50	.75	1.00								
5. Oral Stories	−.23	.29	.06	.19	.22	1.00							
6. Written Stories	−.42	.12	.19	.28	.37	.11	1.00						
7. Cartoon	.03	.08	.20	.27	.38	.15	.24	1.00					
8. $STAT_{Creative}$	1.04	.35	.47	.60	.55	.07	.27	.28	1.00				
9. $STAT_{Practical}$	.46	.25	.43	.57	.53	.14	.32	.29	.61	1.00			
1. $STAT_{Analytical}$	1.50	.24	.43	.62	.53	.12	.22	.22	.57	.62	1.00		
11. Everyday Situational Judgment	−.94	.14	.17	.26	.28	.12	.23	.14	.28	.26	.17	1.00	
12. College Life Questionnaire	−.96	.16	.23	.27	.24	.02	.15	.12	.39	.30	.23	.59	1.00
13. Common Sense Questionnaire	−.95	.27	.24	.28	.26	.20	.31	.04	.38	.32	.26	.55	.33

Nominal $N=777$; FIML used to estimate statistics.
The estimation of correlations in FIML is partially dependent on the variables included in the model. The correlations reported in Table 11 are used for the hierarchical regressions reported in Tables 12 and 13.
[a] z-score transformation applied.

practical ability were added, resulting in a small increase in R^2. Notably, the latent variable representing the common variance among the practical performance measures and high school GPA were the only variables to significantly account for variance in college GPA in Step 3. The inclusion of the creative measures in the final step of this regression indicates that by supplementing the SAT and high school GPA with measures of analytical, practical, and creative abilities a total of 24.8% of the variance in GPA can be accounted for. Inclusion of the triarchic measures in Steps 2, 3, and 4 represents an increase of about 90% (from .159 to .248) in the variance accounted for over and above the typical predictors of college GPA. Table 12.1 and .2 presents the analyses with and without high school GPA (see WebTables 12 A–N for various parallel analyses). The pattern of results was similar even when the SAT and high school GPA variables were entered into the regression equation after the creative, analytic, and practical indicators (Table 13).[20] In sum, across multiple models tested, the triarchic measure added to the prediction by anywhere from 5% to 10.2% (7.4% on average), accounting for up to 50% of the total explained variance in the criterion.

3.8. Group differences

Although one important goal of the present study was to predict success in college, another important goal involved developing measures that reduce socially defined racial and ethnic group differences in mean levels. There are a number of ways one can test for group differences in these measures, each of which involves a test of the size of the effect of race. We chose two: omega square (ω^2), and Cohen's d.

We first considered the omega-square coefficients. This procedure involves conducting a series of one-way analyses of variance (ANOVA) considering differences in mean performance levels among the six ethnic and socially defined racial groups reported, including White, Asian, Pacific Islander, Latino, Black, and Native American, for the following measures: the baseline measures (SAT-V and SAT-M), the STAT ability scales, the creativity performance tasks, and the practical ability performance tasks. The omega-squared coefficient indicates the proportion of variance in the variables that is accounted for by the self-reported ethnicity of the participant. The F-statistic for each ANOVA, its significance, the n on which each analysis was based, and the omega squared for each analysis are presented in Table 14.

The test of effect sizes using the Cohen's d statistic allows one to consider more specifically a standardized representation of specific group differences. The Cohen's d statistic is represented in Table 15. For the test of ethnic group differences, each entry represents how far away from the mean for Whites each group performs in terms of standard deviations. For the test of gender differences, the entries represent how far away women perform from men in terms of standard deviations.

These results indicate two general findings. First, in terms of overall differences represented by omega squared, the triarchic tests appear to reduce race and

[20] A complete set of corresponding tables paralleling WebTables 12 A–N is available on request.

Table 11.2
Actual correlation coefficients computed in SPSS using only complete data with pairwise deletion of missing data

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12
1. College GPA[a]	.00	1.00	756	654	372	370	450	450	186	433	736	477	491	487
2. Everyday Situational Judgment	− .95	.21	.14	671	346	315	425	425	169	368	655	429	439	431
3. College Life Questionnaire	− .96	.22	.10	.60	383	0	226	226	94	222	374	239	245	246
4. Common Sense Questionnaire	− .95	.24	.31	.54	(a)	379	227	227	99	210	369	242	249	245
5. SAT-M[a]	− .04	1.02	.29	.27	.29	.29	457	457	96	256	444	273	310	303
6. SAT-V[a]	− .05	1.03	.27	.29	.25	.26	.75	457	96	256	444	273	310	303
7. Oral Stories	− .18	.97	.25	.07	.08	.08	.19	.21	197	0	188	129	124	127
8. Written Stories	− .41	.97	.12	.24	.10	.32	.30	.40	(a)	441	435	274	284	287
9. Cartoon	.03	.86	.08	.14	.10	.06	.25	.36	.17	.23	757	476	495	485
1. $STAT_{Creative}$	1.03	1.18	.35	.30	.38	.38	.59	.54	.16	.30	.30	490	236	235
11. $STAT_{Practical}$	.43	.97	.24	.30	.32	.37	.56	.53	.15	.30	.33	.65	502	253
12. $STAT_{Analytical}$	1.55	1.31	.25	.13	.18	.25	.61	.52	.03	.24	.17	.53	.62	500
13. High school GPA[a]	.00	1.00	.36	.20	.30	.24	.56	.48	− .01	.27	.20	.45	.46	.47

Correlations coefficients are listed in the bottom triangle. Sample sizes for each correlation coefficient are listed on the top triangle.
[a]Cannot be computed because at least one of the variables is constant.

Table 12
Incremental prediction of college GPA using the triarchic abilities (1) above and beyond the SAT and high school GPA and (2) above and beyond SAT

1.	Step 1	Step 2	Step 3	Step 4
SAT/HSGPA				
Verbal[a]	.098	.084	.066	.005
Math[a]	.070	.011	− .008	− .069
High school GPA[a]	.285∗	.276∗	.267∗	.270∗
Analytical				
Analytical STAT		.096	.054	.012
Practical				
Performance latent[b]			.119∗	.049
Practical STAT			.025	− .033
Creative				
Written				.003
Oral				.273∗
Cartoons				− .072
Creative STAT				.258∗
R^2	.156	.152	.159	.248

2.	Step 1	Step 2	Step 3	Step 4
SAT				
Verbal[a]	.145∗	.125	.098	.039
Math[a]	.188∗	.114	.082	.021
Analytical				
Analytical STAT		.122∗	.068	.021
Practical				
Performance latent[b]			.133∗	.058
Practical STAT			.055	− .015
Creative				
Written				− .003
Oral				.252∗
Cartoons				− .068
Creative STAT				.290∗
R^2	.098	.099	.110	.199

Entries are standardized beta coefficients. *$p<.05$; $N=777$.
[a]z-score transformation applied; [b]see Fig. 2.

ethnicity differences relative to traditional assessments of abilities such as the SAT. Second, in terms of specific differences represented by Cohen's d, it appears that Latino students benefit the most from the reduction of group differences. Black students, too, seem to show a reduction in difference from the White mean for most of the triarchic tests, although a substantial difference appears to be maintained with the practical performance measures. Important reductions in differences can also be seen for Native Americans relative to Whites;

Table 13
Predicting college GPA above and beyond the triarchic abilities using SAT and high school GPA

	Step 1	Step 2	Step 3	Step 4	Step 5
Practical					
Performance latent[b]	.163∗	.378	.058	.058	.049
Practical STAT	.165∗	− .032	− .025	− .015	− .033
Creative					
Written		.018	.020	− .003	.003
Oral		.258∗	.258∗	.252∗	.273∗
Cartoons		− .065	− .069	− .068	− .072
Creative STAT		.356∗	.330∗	.290∗	.257∗
Analytical					
Analytical STAT			.026	.021	.012
SAT					
Verbal[a]				.039	.005
Math[a]				.021	− .069
HSGPA					
High school GPA[a]					.270∗
R^2	.075	.208	.201	.199	.248

Entries are standardized beta coefficients. *$p<.05$; $N=777$.
[a]z-score transformation applied; [b]see Fig. 2.

Table 14
Amount of variance in each assessment accounted for by ethnicity, using the omega-square effect size statistic

Measure	*F*	*p*	*N*	Omega squared (ω^2)
SAT				
Verbal	35.8	<.001	341	.09
Math	15.2	<.001	341	.04
Total (combined)	28.2	<.001	340	.07
STAT				
Analytical	0.5	ns	370	.00
Practical	12.8	<.001	374	.03
Creative	6.7	<.01	369	.02
Practical performance				
EDSJ (Movies)	5.9	<.05	493	.01
Common Sense	2.6	ns	273	.01
College Life	8.4	<.01	298	.02
Creative performance				
Cartoon captions	14.0	<.001	569	.02
Oral Stories	6.0	<.05	152	.03
Written Stories	3.1	ns	329	.01

however, the very small sample size suggests that any conclusions about Native American performance should be made tentatively. In addition, mean differences between groups are a function of score reliabilities (Table 15). Although the reliabilities for the Rainbow tests are adequate (see discussion above), the brevity of the tests decreases their values somewhat. Correspondingly, the levels of reliabilities of the Rainbow scores should be kept in mind and these group results should be interpreted with caution.

Although the group differences are not perfectly reduced, these findings suggest that measures can be designed that reduce ethnic and socially defined racial group differences on standardized tests, particularly for historically disadvantaged groups such as Blacks and Latinos. These findings have important implications for reducing adverse impact in college admissions.

4. Discussion

4.1. Summary of findings

The SAT is based on a conventional psychometric notion of cognitive skills. Based on this notion, it has had substantial success in predicting college performance. But perhaps the time has come to move beyond conventional theories of cognitive skills. Based on multiple regression analyses, for our sample, the triarchic measures alone approximately double the predicted amount of variance in college GPA when compared with the SAT alone (comparative R^2 values of .199 to .098, respectively). Moreover, the triarchic measures predict an additional 8.9% of college GPA beyond the initial 15.6% contributed by the SAT and high school GPA. These findings, combined with the substantial reduction of between-ethnicity differences, make a compelling case for furthering the study of the measurement of analytical, creative, and practical skills for predicting success in college.

4.1.1. Analytical skills: SAT, HSGPA, $STAT_{Analytical}$

It is not surprising to find that analytical skills as tapped by the SAT, high school GPA, and the $STAT_{Analytical}$, are important to successful performance in college. And it is not altogether surprising that high

Table 15
Group differences as represented by the Cohen's *d* statistic

	Test reliability	Black *d*	(*N*)	Latino *d*	(*N*)	Asian *d*	(*N*)	Native Am. *d*	(*N*)	Women *d*	(*N*)
SAT measures											
Math		−.74	31	−.98	53	.35	48	−1.00	4	−.37	264
Verbal		−.67	31	−1.10	53	−.23	48	−.62	4	−.17	264
Total (combined)		−.73	31	−1.10	53	.04	47	−.76	4	−.28	266
STAT											
Analytical	0.59	−.19	31	−.36	55	.34	51	−.33	8	−.30	290
Practical	0.53	−.47	31	−.53	53	.09	55	−.66	7	−.18	297
Creative	0.60	−.67	31	−.46	61	−.03	55	−1.15	7	−.18	288
Practical performance											
ESDJ (Movies)	0.76	−.51	46	−.35	77	.05	82	−.77	4	.19	384
Common sense	0.91	−.89	15	−.22	44	.21	37	−.40	8	.52	222
College life	0.89	−.68	32	−.22	41	−.22	50	.20	3	−.05	229
Creative performance											
Cartoon captions	0.86	−.24	45	−.51	86	−.16	85	−.39	10	−.13	443
Oral Stories	0.79	−.14	16	−.46	27	−.50	25	.50	2	.04	111
Written Stories	0.80	−.26	21	−.11	51	−.25	45	.01	7	.00	269

For ethnicity, Whites are the reference group. For gender, men are the reference group.

school GPA turns out to be the best predictor, subsuming the unique contribution of other largely analytical measures and including other variance as well.

High school GPA is such a good predictor because the best predictor of future behavior of a certain kind is past behavior of the same kind. The best predictor of grades should be–and generally is–grades. GPA is psychologically complex. It thus is not of much use as a *psychological* predictor because it contains so many things confounded within it, including conscientiousness.

We concur that analytical abilities are necessary for success in many academic disciplines. However, we have also suggested that these abilities may not be sufficient for college success—particularly for disadvantaged students. Although the zero-order correlation between the $STAT_{Analytical}$ test and GPA was approximately the same magnitude as the correlation between the SAT and GPA, the analytical section of our test added little in terms of predictive power over and above the SAT. Because the SAT is already a well developed, reliable, and valid measure of analytical skills, we plan to dispense with the analytical section in future versions of the Rainbow Project and simply use the SAT as our analytical measure.

Given that traditional pedagogy emphasizes memory and analytical skills, it may not be particularly clear what other student characteristics might determine success. The theory of successful intelligence proposes that creative and practical abilities are also important for success in many areas of life, including college; for example, creative abilities are important in creating course projects, written essays, and papers, and practical skills are important in understanding how to study for exams, manage time, and infer professors' expectations for coursework.

4.1.2. Creative skills: Cartoons; Oral Stories; Written Stories; $STAT_{Creative}$

The creative performance measures provide modest reliability and zero-order prediction of college GPA. Research in creativity has repeatedly demonstrated the multidimensional characteristic of this construct (see, for example, Sternberg, 1999b), and our analyses suggest that our measures of creativity show similar multidimensionality. Our measures do show some common variation with verbal skills; however, there is evidence to suggest that reliable variation in the Oral Stories is being determined by skills distinct from the traditional academic abilities assessed by the SAT-V. When incremental prediction is considered, both the Oral Stories task and the $STAT_{Creative}$ remain significant predictors of academic performance (college GPA) beyond SAT.

4.1.3. Practical skills: movies, common sense, college life, $STAT_{Practical}$

The failure of the $STAT_{Practical}$ to load on the practical measure suggests that there are very strong method factors. Practical skills cannot be fully measured by the kinds of multiple-choice measures that appear on the original STAT. For this reason, in the new Rainbow measures, creative and practical items will be mostly performance-based.

The practical performance measures have good reliability and appear to be effective measures of tacit knowledge and practical skill. Together, the three practical performance tests load on a higher-order practical-skill factor, although they did not significantly predict college GPA at the .05 alpha level after the creative measures were entered into the regression equation.

There are a number of issues yet to be resolved, including deciding on an appropriate criterion against which to assess responses to the practical measures, and whether it should be sample-based or expert-based. Another issue with practical intelligence is that cumulative high school GPA is likely to reflect some and perhaps many of the practical skills necessary for academic success, particularly because it reflects academic success over an extended period of time and not simply in a single testing situation.

4.1.4. Incremental predictive power

Overall, in the full regression, the only analytical indicator that provided statistically significant prediction of college GPA was high school GPA. The practical performance measures did load on a common factor, and the practical variables were statistically significant predictors of college GPA when they were the only variables in the model. However, when the full range of analytical, practical, and creative measures were added to the regression prediction equation, the latent practical factor was not a statistically significant predictor of college GPA. With regard to the creative measures, two of the four indicators (i.e., the $STAT_{Creative}$, the Oral Stories, and the Cartoons) were statistically significant predictors in the final regression equation, even after including SAT and high school GPA variables, as well as practical indicators. Yet, we need to issue a note of caution here, hoping that others in the field or our future studies will explore certain unexpected results.

For example, the Cartoons variable demonstrated a negative regression weight, indicating, in combination with a relatively low correlation with GPA, the likely presence of negative net suppression in the regression (Krus & Wilkinson, 1986). In attempt to explain this connection, we suggest that the Cartoons indicator might also capture and reflect the trait of humor. Humor, in its different aspects, might positively relate to creativity, but might also negatively relate to academic success. In theoretical models of humor, there is typically a facet of challenge to authority and disobedience (Barron, 1999). Clearly, these aspects of humor, when demonstrated in the classroom, might not positively influence college, especially freshman, GPA. Because we do not have data to explain this effect completely, we acknowledge this as a limitation of our study. Yet, suppression effects are rather common in complex models and are not viewed as a flaw of a measure, design, or model (Cohen, Cohen, West, & Aiken, 2003). In our case, further research is needed to explain this finding completely.

4.1.5. Group differences

One important goal for the current study, and future studies, is the creation of standardized test measures that reduce the different outcomes between different groups as much as possible in a way that still maintains test validity. Our measures suggest positive results toward this end. Although the group differences in the tests were not reduced to zero, the tests did substantially attenuate group differences relative to other measures such as the SAT. This finding could be an important step toward ultimately ensuring fair and equal treatment for members of diverse groups in the academic domain.

There has been a lot of buzz in the psychological literature about relations of ability-test scores to socially defined ethnic group membership. People have widely differing views on score differences and what they mean (e.g., Rushton & Jensen, 2005, versus Sternberg, 2005; see also Hunt & Sternberg, in press). We do not wish to contribute to the noise level of this debate. Our results suggest, however, that there may be variation relevant to college performance that is not tapped by conventional tests.

4.2. Methodological issues

Although this first study presents a promising start for the investigation of an equitable yet powerful predictor of success in college, the study is not without its share of methodological problems. Future development of these tests will help sort out some of the problems borne out of the present findings.

4.2.1. Problems with the sample

At this stage of the project, our goal was to recruit a broad range of higher education institutions to participate in the investigation. Participating institutions included community colleges, 4-year colleges, and universities. It is important to note, however, that the participants in this project reflect a purposive sample rather than a truly random sample of higher education institutions.

4.2.2. Problems with the creativity tests

One important problem raised by the creativity tests is that they risk tapping into verbal skills too much. The structural equation model suggests a very strong path between the creativity latent variable and SAT-V (0.78); however, the correlations between the SAT-V test and the individual creativity performance measures (Cartoons, Oral and Written Stories) were consistently less than .40. This finding, along with the fact that creativity measures do predict college GPA above and beyond the SAT, suggests that there is more to the creativity tests than mere verbal-expression skills. Moreover, the best predictor was Oral, not Written Stories, and Oral Stories require less sophisticated verbal and especially lexical skills than do written ones.

But other research has shown that verbal skills may play an important role in creativity anyhow. Measures of "verbal fluency" have had a long history in the conceptualization of measures of creativity, from Guilford (1967) to Mednick and Mednick (1967) to Torrance (1974). Recent empirical evidence shows that raters on written story products often have difficulty removing quality of written expression from their judgments of creativity (Sternberg et al., 1996). Although verbal fluency can be distinguished from strict verbal comprehension (Carroll, 1993; Sincoff & Sternberg, 1987; Thurstone, 1938), one should not expect the relationship between creativity and verbal skills to be completely orthogonal.

Why might SAT-V correlate with our verbal creativity measures? For one thing, there may be content effects due to shared verbal representations. Both sets of tests presumably access the same lexical mental representations. Better scores would be associated with richer and more interconnected representations. In addition, retrieval processes may be shared. In both cases, the participant needs to access the representations, and the retrieval processes may be related or identical

for the two tasks. But the two tasks are not the same, in that one requires primarily verbal comprehension, and the other, verbal production (cf., e.g., Thustone's (1938) distinction between verbal comprehension and verbal fluency).

Nevertheless, future studies could encourage raters to discriminate more between verbal ability and creative skills in their judgments of creativity of stories and captions. For example, we could simplify the ratings by using only two or three dimensions for judgment per task, and having one of those dimensions directly involve a judgment of verbal expression. This system might help future judges distinguish between verbal ability and creativity, much the same way our judges were able to distinguish between task appropriateness and indicators of creativity on the Cartoon task. Including this rating could allow judges to more easily recognize a well-written but not particularly creative story, or a poorly written but highly creative story, sharpening our measurement of creativity.

Future studies should also allow for more than one judge for all written and oral stories. We were able to demonstrate unique predictive power for our creative measures that involved one rater for many stories; however, no single rater can be a perfect judge of creativity. In conjunction with more refined rubrics, having multiple raters could allow for a more accurate, and more powerful, measure of the creativity construct.

4.2.3. Problems with the practical tests

Although the tests of practical skills did show zero-order relationships to college GPA, these relationships were reduced to marginal significance in the context of a multiple regression that included all variables. One interpretation of this finding is that our operationalization of practical skills may not fully capture the kinds of practical skills necessary for success in college. For example, the different types of scenarios or response options might possibly better capture the kinds of practical skills necessary for success in college.

A second possibility is that the effects of practical skills could be reduced because of the use of the sample-based profile against which our participants' scores were determined. Using an expert panel (e.g., scores of already successful undergraduates) might provide a more refined profile for determining participants' scores, and could reduce the possibility that our measures are merely capturing conformity to peers. However, there are problems that arise from using expert panels, as discussed earlier. Expert panels show a great deal of variation in terms of appropriate responses on measures of practical skills; they might introduce cultural or group biases depending on the composition of the panel, and they might not contribute a more powerful profile of practical intelligence than would a sample-based profile.

One might argue that the failure of the STAT multiple-choice tests to provide independent prediction is an embarrassment to the theory of successful intelligence. We disagree. Galton's (1883) initial measurements of simple constructs were not very successful. For many years, people concluded, wrongly, that all such measures were invalid. Eventually, in the 1970s, investigators such as Hunt et al. (1973) showed that this thinking was wrong. The early measures were just too crude. In the same way, our initial multiple-choice measures of practical skills did not prove to be psychometrically distinct from *g*. Had we never found any measures that were psychometrically distinct from *g*, that would have been a problem. But we have (see also Sternberg et al., 2000). So our data suggest that the problem was with the initial measures, not the theory. Over time, still better measures perhaps will be created.

4.2.4. Problems with test length

The test as constructed was too long to administer in its entirety to our participants. Hence, we used the incomplete design and short versions of the assessments. This was an exploratory study to suggest ways of cutting down the length of the test to the 2-h maximum that will be imposed in any future version. Inevitably, this situation impacted the reliabilities of our measures and our ability to manipulate complete data.

5. Conclusion

The theory of successful intelligence appears to provide a strong theoretical basis for augmented assessment of the skills needed for college success. There is evidence to indicate that it has good incremental predictive power, and serves to increase equity. As teaching improves and college teachers emphasize further the creative and practical skills needed for success in school and life, the predictive power of the test may increase. Cosmetic changes in testing over the last century have not made great differences to the construct validity of assessment procedures. The theory of successful intelligence could provide a new opportunity to increase construct validity. We are not suggesting that this theory is unique in providing such opportunities. But we do believe that, given the data, its value in creating a new generation of assessments for future use in college admissions is at least worthy of exploration.

Acknowledgments

The research reported in this article was supported by the College Board. The authors wish to thank Ms. Robyn Rissman for her editorial assistance. Correspondence concerning the article should be sent to Robert J. Sternberg, Office of the Dean of Arts and Sciences, Ballou Hall 3rd Floor, Tufts University, Medford, MA 02155.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.intell.2006.01.002.

References

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and noncognitive determinant and consequences of complex skill acquisition. *Journal of Experimental Psychology. Applied*, *6*, 259−290.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1−23.

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, *17*, 71−103.

Barron, J. W (Ed.). (1999). *Humor and psyche: Psychoanalytic perspectives* Hillsdale, NJ, US: Analytic Press, Inc.

Binet, A., & Simon, T. (1916). *The development of intelligence in children.* Baltimore: Williams and Wilkins Originally published in 1905.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Boring, E. G. (1923, June 6). Intelligence as the tests test it. *New Republic*, 35−37.

Bridgeman, B. (2004). Unbelievable results when predicting IQ from SAT scores. *Psychological Science*, *16*, 745−746.

Bridgeman, B., Burton, N., & Cline, F. (2001). *Substituting SAT II: Subject tests for SAT I: Reasoning test: Impact on admitted class composition and quality.* New York: College Entrance Examination Board College Board Report No. 2001−3.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test.* New York: College Entrance Examination Board College Board Report No. 2000–1.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new structure of intellect. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 27−56). Hillsdale, NJ: Erlbaum.

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

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

Cianciolo, A. T., & Sternberg, R. J. (2004). *A brief history of intelligence.* Malden, MA: Blackwell.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Deary, I. J. (2000). *Looking down on human intelligence.* Oxford, UK: Oxford University Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society*, *39*, 1−38.

Detterman, D. K. (1986). Human intelligence is a complex system of separate processes. In R. J. Sternberg, & D. K. Detterman (Eds.), *What is intelligence?* (pp. 57−61). Norwood, NJ: Ablex.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores*. College Board Rep. No. 99-1; ETS Rep. RR. No. 99-2.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology. General*, *128*, 309−331.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? *Psychological Science*, *15*, 373−378.

Frey, M. C., & Detterman, D. K. (2005). Regression basics. *Psychological Science*, *16*, 747.

Galton, F. (1883). *Inquiry into human faculty and its development.* London: Macmillan.

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

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century.* New York: Basicbooks.

Glass, G. V., & Hopkins, K. H. (1996). *Statistical methods in education and psychology.* Boston: Allyn and Bacon.

Greenfield, P. M. (1997). You can't take it with you: Why abilities assessments don't cross cultures. *American Psychologist*, *52*, 1115−1124.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P. (1982). Cognitive Psychology's ambiguities: Some suggested remedies. *Psychological Review*, *89*, 48−59.

Hezlett, S., Kuncel, N., Vey, A., Ones, D., Campbell, J., & Camara, W. J. (2001). *The effectiveness of the SAT in predicting success early and late in college: A comprehensive meta-analysis.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Hunt, E. B. (1980). Intelligence as an information-processing concept. *British Journal of Psychology*, *71*, 449−474.

Hunt, E. B. (1995). *Will we be smart enough? A cognitive analysis of the coming workforce.* New York: Russell Sage Foundation.

Hunt, E., Frost, N., & Lunneborg, C. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, *vol. 7* (pp. 87–122). New York: Academic Press.

Hunt, E. B., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology*, *7*, 194−227.

Hunt, E., & Sternberg, R. J. (in press). Sorry, wrong numbers: An analysis of a study of a correlation between skin color and IQ. *Intelligence.*

Intelligence and its measurement: A symposium, (1921). *Journal of Educational Psychology, 12,* 123–147, 195–216, 271–275.

Jensen, A. R. (1998). *The g factor.* Westport: Praeger-Greenwood.

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the Nation.* New York: College Entrance Examination Board College Board Report No. 2002–6.

Krus, D. J., & Wilkinson, S. M. (1986). Demonstration of properties of a suppressor variable. *Behavior Research Methods, Instruments, and Computers*, *18*, 21−24.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence*, *14*, 389−433.

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence*, *21*, 247−266.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze, & R. D. Roberts (Eds.), *International handbook of emotional intelligence.* Berlin, Germany: Hogrefe and Huber.

Linacre, J. M. (1989). *FACETS user's guide.* Chicago: MESA Press.

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

Linacre, J. M. (1998). *Many-facet Rasch measurement* [on-line]. Available: http://www.winsteps.com/facetman/index.htm

Lubart, T. I., & Sternberg, R. J. (1995). An investment approach to creativity: Theory and data. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 269−302). Cambridge, MA: MIT Press.

Mackintosh, N. J. (1998). *IQ and human intelligence.* Oxford: Oxford University Press.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT v2.0. *Emotion*, *3*(1), 97−105.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, *29*(4), 409−454.

McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research*, *18*(3), 145−166.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braveman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730−740.

Mednick, S. A., & Mednick, M. T. (1967). *Remote associates test examiner's manual.* Boston: Houghton-Mifflin.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel, & G. R. Wheaton (Eds.), *Applied measurement methods in Industrial Psychology* Palo Alto, CA: Davies-Black Publishing.

Muthen, L. K., & Muthen, B. O. (2002). *Mplus user's guide* (Second edition). Los Angeles, CA.

Neubauer, A. C., & Fink, A. (2005). Basic information processing and the psychophysiology of intelligence. In R. J. Sternberg, & J. E. Pretz (Eds.), *Cognition and intelligence* (pp. 68−87). New York: Cambridge University Press.

Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.

Peterson, C., & Seligman, M. E. (2004). *Character strengths and virtues.* New York: Oxford University Press.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language and ethnic groups.* New York: College Entrance Examination Board College Board Report No. 93-1, ETS RR No. 94-27.

Rencher, A. C. (1995). *Methods of multivariate analysis.* New York: Wiley and Sons.

Rogers, W. A., Hertzog, C., & Fisk, A. D. (2000). An individual differences analysis of ability and strategy influences: Age-related differences in associative learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(2), 359−394.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235−294.

Schmidt, K. M., Bowles, R. P., Kline, T. L., & Deboeck, P. (March, 2002). *Psychometric Scaling Progress Report: The Rainbow Project Data — revised.* Technical report presented to the College Board.

Sincoff, J., & Sternberg, R. J. (1987). Two faces of verbal ability. *Intelligence*, *11*, 263−276.

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *American Journal of Psychology*, *15*(2), 201−293.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1980). Sketch of a componential subtheory of human intelligence. *Behavioral and Brain Sciences*, *3*, 573−584.

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. (1990). *Metaphors of mind.* New York: Cambridge University Press.

Sternberg, R. J. (1993). *Sternberg Triarchic Abilities Test.* Unpublished test.

Sternberg, R. J. (1997). *Successful intelligence.* New York: Plume.

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

Sternberg, R. J. (Ed.) (1999b). *Handbook of creativity.* New York: Cambridge University Press.

Sternberg, R. J. (Ed.) (2000). *Handbook of intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (2003). *Wisdom, intelligence, and creativity, synthesized.* New York: Cambridge University Press.

Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist*, *59*(5), 325−338.

Sternberg, R. J. (2005). There are no public-policy implications: A reply to Rushton and Jensen. *Psychology, Public Policy, and Law*, *11*, 295−301.

Sternberg, R. J., & Clinkenbeard, P. R. (1995). A triarchic model applied to identifying, teaching, and assessing gifted children. *Roeper Review*, *17*(4), 255−260.

Sternberg, R. J., & Detterman, D. K. (1986). *What is intelligence?* Norwood, N.J.: Ablex Publishing Corporation.

Sternberg, R. J., Ferrari, M., Clinkenbeard, P. R., & Grigorenko, E. L. (1996). Identification, instruction, and assessment of gifted children: A construct validation of a triarchic model. *Gifted Child Quarterly*, *40*, 129−137.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., et al. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., & Kidd, K. K. (2005). Intelligence, race, and genetics. *American Psychologist*, *60*, 46−59.

Sternberg, R. J., Grigorenko, E. L., & Singer, J. L. (Ed.) (2004). *Creativity: The psychology of creative potential and realization* Washington: American Psychological Association.

Sternberg, R. J., & Hedlund, J. (2002). Practical intelligence, *g*, and work psychology. *Human Performance*, *15*(1/2), 143−160.

Sternberg, R. J., Lautrey, J., & Lubart, T. I. (Ed.) (2003). *Models of intelligence for the new millennium* Washington, DC: American Psychological Association.

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

Sternberg, R. J., & Pretz, J. E. (Ed.) (2005). *Cognition and intelligence* New York: Cambridge University Press.

Sternberg and Project Rainbow Collaborators. (2003 revision). *The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills*. Technical report submitted to the College Board.

Sternberg, R. J., & Wagner, R. K. (1993). The *g*-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, *2*(1), 1−4.

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

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist*, *50*(11), 912−927.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Torrance, E. P. (1974). *Torrance tests of creative thinking: Norms—technical manual.* Lexington, MA: Ginn.

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

Wagner, R. K., & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. J. Sternberg, & R. K. Wagner (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world* (pp. 51−83). New York: Cambridge University Press.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore: Williams and Wilkins.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (Eds.) (1990). *Predicting college grades: An analysis of institutional trends over two decades* Princeton, NJ: Educational Testing Service.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. Little, K. Schnabel, & J. Baumert (Eds.), *Modeling Longitudinal and Multilevel Data*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Wu, M., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: User's manual.* Australia: ACER.