Journal of Educational Measurement
Spring 2002, Vol. 39, No. 1, pp. 1–37

Grades and Test Scores: Accounting for Observed Differences

Warren W. Willingham
Judith M. Pollack
Charles Lewis
Educational Testing Service

Why do grades and test scores often differ? A framework of possible differences is proposed in this article. An approximation of the framework was tested with data on 8,454 high school seniors from the National Education Longitudinal Study. Individual and group differences in grade versus test performance were substantially reduced by focusing the two measures on similar academic subjects, correcting for grading variations and unreliability, and adding teacher ratings and other information about students. Concurrent prediction of high school average was thus increased from 0.62 to 0.90; differential prediction in eight subgroups was reduced to 0.02 letter-grades. Grading variation was a major source of discrepancy between grades and test scores. Other major sources were teacher ratings and Scholastic Engagement, a promising organizing principle for understanding student achievement. Engagement was defined by three types of observable behavior: employing school skills, demonstrating initiative, and avoiding competing activities. While groups varied in average achievement, group performance was generally similar on grades and tests. Major factors in achievement were similarly constituted and similarly related from group to group. Differences between grades and tests give these measures complementary strengths in high-stakes assessment. If artifactual differences between the two measures are not corrected, common statistical estimates of validity and fairness are unduly conservative.

Many of the most important educational decisions we make about young people concern the summative, often irreversible, judgments regarding entry to or exit from programs or institutions. Who will be placed in a slow or fast track in grade school, earn a high school diploma, be accepted in a selective college, or be admitted to a demanding graduate or professional program? Grades and test scores are the two types of evidence most commonly used in supporting these judgments.

When a cumulative grade record is used in reaching an important educational decision, it becomes, in effect, a high-stakes predictor or criterion. In this capacity grades take on a broader assessment function that is different from the teachers' original evaluations of their students' acquired proficiency in a particular subject in a given class. Tests are routinely evaluated for the broader function, grades less systematically. In serving the broader purpose, grade averages have virtues as well as limitations. Understanding these characteristics of grades is important for the valid use of test scores as well as grade averages because, in practice, the two measures are often intimately connected.

Note the paradox: We use tests to keep grade scales honest or because we do not fully understand or trust grades to be an accurate indicator of educational outcomes;

yet, we use grades to demonstrate the validity and fairness of tests and to justify their use. One likely source of the apparent contradiction is the tendency to assume that a grade average and a test score are, in some sense, mutual surrogates; that is, measuring much the same thing, even in the face of obvious differences. One manifestation of that implicit assumption is the common inclination to treat an improvement in grade prediction as the dominant, if not the sole, basis for validating a high-stakes admissions test and justifying its use. Another telling instance lies in the formal professional definition of a biased test; that is, a test that predicts a mean criterion score that is different from the actual mean (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999).

It would be useful to have a better understanding of the relationship between grades and test scores. Insuring the validity and fairness of one requires an appreciation of issues concerning the validity and fairness of the other. Testing and grading have a long history of public controversy (Cronbach, 1975; Cureton, 1971). National agencies and special commissions periodically undertake studies of requirements regarding the technical quality and conditions for proper use of measures used in high-stakes decisions, tests in particular (Gifford & O'Connor, 1992; Heubert & Hauser, 1999; Office of Civil Rights, 1999). Many critics — both inside and outside the profession — have discussed the shortcomings and the social concerns that testing engages (Lemann, 1999; Shepard, 1992b).

Nonetheless, objective measures of school achievement have obvious benefits — especially for high-stakes selection (Beatty, Greenwood, & Linn, 1999, pp. 20–22) and fostering educational accountability and reform (Heubert & Hauser, 1999, pp. 33–40). This is not a new development. Linn (2000) described the use of tests as key elements in five waves of educational reform during the past 50 years. Current goals in assessment reform include establishing better linkages between instructional objectives and testing (Frederiksen & Collins, 1989; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Resnick & Resnick, 1992), focusing assessment on established educational standards (Baker & Linn, 1997; Shepard, 2000), and broadening the range of assessment formats and the skills thus engaged (Bennett & Ward, 1993; Wiggins, 1989). These initiatives imply dissatisfaction with both grading and testing — and the need to better understand and realize the strengths of each.

Our premise was that it should be possible to account for much of the differences observed between grades and test scores. This study had several purposes: to suggest a framework that might help to explain those differences, to evaluate an approximation of the framework, to test its generality in different groups of students, and to examine possible implications of the findings as to the merits of grades and scores in high-stakes decisions.

Differences in Grades and Test Scores

Because we ask why individual students or groups of students often perform somewhat differently on grades and test scores, we are concerned with factors that bear on the selective function. That is, what factors cause grades and test scores to rank students differently and therefore identify somewhat different high or low

scorers. Thus, the present analysis is based on patterns of individual and group differences in assessment outcomes rather than content differences in the measures. It is important to bear in mind, however, that additional features of grades and tests, not considered here, will affect the educational relevance of the measures and the effects of their use.

Grade–test score differences that do not necessarily cause any discrepancy in rank order are not addressed. A trivial instance is discrepancy due to arbitrary scale differences (e.g., a 4.0 grade scale and test scores that range from 100 to 200). A critical instance could be differential difficulty reflected in the passing standard (e.g., a state assessment might fail many more students than the prevailing high school grade scale; see Steinberg, 2000).

For many years researchers have sought to understand what factors are associated with student achievement. The question readily lends itself to causal modes of thinking; namely, what family circumstances and values promote achievement in school and what student characteristics, habits, and attitudes lead to good grades? In an analysis of student development or a prediction of future grades, a common purpose is to "explain," in a statistical sense, what accounts for a student's achievement above or below that expected on the basis of test scores.

Causal logic is not the most useful way to view the relationship between test scores and grades. Test scores do not cause individual differences in grades, nor vice versa. If a test and a grade are intended to represent much the same achievement, then individual differences in both measures presumably result from much the same learning process, influenced by much the same environmental and genetic factors, and channeled by much the same cognitive differences and personal interests. From that perspective, understanding the wellsprings of achievement will not necessarily help to understand differences in grades and test scores.

For the purposes of this study, we pose a somewhat different question, "How does the composition of grades and test scores differ and what are the implications of those differences?" The two measures are presumably somewhat similar composites of skill and knowledge that are generally relevant to the achievement construct of interest plus some other sources of construct-irrelevant variance. From this view, grades and test scores are correlated only moderately because the elements of the two composites overlap only partly. The rank orders of students based on the two measures are different because the different components interact with learning circumstances and individual differences in behavior and background. Table 1 suggests a framework of possible sources of difference between grades and test scores.

If grades and test scores are expected to rank students similarly, perhaps the most obvious assumption is that the two measures encompass knowledge and skills based on a similar academic domain as implied by Category A.1 in Table 1. Some systematic differences are likely. Even if a test and a grade average are based on the same coursework, external tests cannot cover material in the same detail. Teachers and test developers may be inclined to assess somewhat different subject matter. Some teachers may be performance oriented in their testing and grading; that is, stress writing or oral presentation. Thus, like subject matter, assessment format can

TABLE 1
A Framework of Possible Sources of Discrepancy Between Observed Grades and Test Scores

A. Content Differences Between Grades and Test Scores
   1. Domain of general knowledge and skill
      a. Subjects covered, such as science and history; broad divisions within subjects, such as physics or European history
      b. General cognitive skills, such as reasoning, writing, or performance
   2. Specific knowledge and skills as reflected in
      a. Course-based content throughout the school district, state, or nation (especially relevant to an external test)
      b. Classroom-based content (especially relevant to a teacher's grade)
      c. Iindividualized content (especially relevant to personal interests, skills, and course of study)
   3. Components other than subject knowledge and skills
      a. Social objectives of education (e.g., leadership, citizenship)
      b. Academic and personal development (e.g., attendance and participation, completing assignments, disruptive behavior, effort and coping skills, interpersonal competence)
      c. Assessment skills and debilities (pertinent to test-taking or class assignments, general or specific to particular assessments, construct relevant or irrelevant, confidence or anxiety)

B. Individual Differences That Interact With Content Differences
   1. Early development and relevant learning acquired outside of school
   2. Student characteristics that can affect academic motivation
      a. Behavior in and out of school
      b. Attitudes about school and learning
      c. Family circumstances
   3. Teacher judgment regarding the student's performance

C. Situational Differences
   1. Differences across contexts
   2. Differences over time

D. Errors in Grades or Test Scores
   1. Systematic error—Noncomparability
      a. Variation in grading standards (across schools, courses, teachers, and sections)
      b. Variation in test score scales (across forms; across time)
      c. Cheating (by students or schools, on class assignments or tests)
   2. Unsystematic measurement error—Unreliability (in grades and in test scores)

be associated with relevant or irrelevant skills that differ somewhat between grades and external test scores.

Category A.2 suggests an inherent distinction between a teacher's grade and an external test score that stems from different purposes of the two measures. To aid learning and instruction, grades reflect specific knowledge and skills stressed in the particular classes that a given student takes. To foster accountability, standardized tests provide outcome assessments that are comparable across schools (Shepard, 2000). This distinction will result in different performance on grades and test scores because students experience somewhat different curricular and learning situations.

A student may know a good deal about the subject, but if poorly motivated to study the material assigned, he or she is less likely to correctly answer the teacher's specific questions about that particular material and will be graded accordingly.

An external test in a given subject area purports to represent content of a typical course in the subject, wherever it is taught. A student might receive a good score on such a course-based test due to having acquired relevant knowledge and skills outside of school or some years earlier. Other students may have a mediocre command of the subject but earn a good grade in class by working hard on the specific material and exercises presented by their teacher.

Category A.2.c in Table 1 represents another potentially critical difference between grades and test scores. Educators define common learning goals and standards, but students vary in their academic interests and pursuits, and the development of each student is an important goal of education. To the degree that education is individualized, a standardized test will tend to yield results somewhat different from an individualized grade assessment (Keeton & Associates, 1976). Furthermore, students can achieve common educational goals in different ways.

Category A.3 refers to other elements that may be represented in grading or test scores, but are not formally part of the knowledge and skills that define the subject domain. In particular, students may receive grade credits or deductions for particular behavior such as attendance, class participation (or disruption), and turning in homework assignments (A.3.b). These elements can influence the students' grade even if they have little effect on subject mastery. Their construct relevance in grading stems from a broader definition of education that includes conative aspects of learning like volition and effort. Construct irrelevant elements of the assessment process can affect both grades and test scores positively or negatively. Examples include test wiseness and anxiety in testing or performing situations.

Category B refers to individual differences that interact with the content differences of Category A, thereby creating grade–test score differences. The extent to which students do better on class-based grades than on course-based tests will tend to depend on their total learning experiences (B.1), their dedication to schoolwork (B.2), and the teacher's judgment (including positive halo and negative bias) of how well they have performed in class (B.3).

Category C recognizes that grade–test score differences can result from situational variation in the student's motivation or the character of the learning experience. If grades and test scores are based upon learning at different ages and types of institutions (e.g., high school and college), there is higher likelihood of content differences and an added level of interaction between Category A and Category B. Thus, grades and tests should be based on concurrent learning in a similar context if they are expected to reflect a similar rank order in performance.

In Category D, two types of error cause discrepancies between observed grades and test scores: systematic and unsystematic. Much research suggests that noncomparability is an important form of systematic error in grades (D.1.a). A similar problem with test scores (D.1.b) can occur if, for example, a practitioner uses percentile scores based on the wrong norm group or makes a consequential equating error.

Another type of noncomparability is the possibility of a change in test difficulty over time due to increased familiarity of some students following a particular test form's repeated use (Cannell, 1988). Noncomparability can result from cheating (D.1.c), evidently a common student practice, but often notorious when it involves a high-stakes test. Schools also engage in questionable practices involving both tests and grades (Saslow, 1989; Wilgoren, 2000). Unreliability is, by definition, unsystematic measurement error (D.2) that affects both grades and test scores, independent of noncomparability.

A Five-Factor Approximation

It is not feasible to evaluate empirically all aspects of Table 1, but since the elements overlap, it is unnecessary to cover the full framework to account for a significant proportion of observed grade–test score differences. Choosing factors for an approximation of Figure 1 depends on judging which elements are likely to be most important and which can be controlled with a good database. High school seniors from the National Education Longitudinal Study of 1988 (NELS) provided excellent data for our purpose. Using this database partly controlled situational variation (Category C) and offered extensive data relevant to the other categories. Our analysis was based on these five factors: (a) Subjects Covered, (b) Grading Variations, (c) Reliability, (d) Student Characteristics, and (e) Teacher Ratings.

Before describing these factors, it is useful to comment briefly on the analytic model. Principal aims of this study were to determine to what extent the proposed factors can account for differences in students' rank order of their high school grade average and their NELS Test score and to evaluate what role each factor plays. This is why the study examines differences between grades and test scores by analyzing patterns of individual and group differences rather than analyzing content or structural differences between the two measures.

The five-factor approximation derives from the assumption that grades and test scores would correspond more closely — that is, the measures would be more highly correlated — if one could alter the components or statistically adjust the two measures so that they are similarly constituted. Prediction is a useful analytic framework because we can ask, simply, "How might one add to or correct one measure to account for variation in the other?"

There are three means of doing so. The most obvious approach is to develop an index; that is, give each student a score representing the factor and include it as a variable in the analysis. Another method is to define or correct the grade or test score so as to reduce apparent differences in their constitution. Finally, some factors can be handled as statistical adjustments. Thus, representing each factor in the analysis involves somewhat different steps and assumptions. Also, Factors 1, 2, and 3 are different from Factors 4 and 5. The former three concern mismatched material and error components that make grades and test scores less comparable. The latter two concern student behavior and other characteristics that were postulated to interact differently with grade performance and test performance.

Factor 1: Subjects Covered

Making certain that a grade and a test score are based on generally similar subject matter pertains mainly to Category A.1 in Table 1. The NELS survey provides a complete transcript for each student including traditional academic subjects, vocational subjects, and service courses like driver training and physical education. The NELS tests cover a more restricted range of academic subjects: reading, mathematics, science, and social studies. A reasonably good match can be attained in two steps: (a) restricting a High School Average (HSA) to courses in English, mathematics, science, and social studies, the four "New Basics" subjects (Ingels et al., 1995) that correspond most closely to the four NELS tests; and (b) weighting the four tests to form a NELS Composite that best reproduces the students' rank order on HSA.

Factor 2: Grading Variations

Variation in grading standards (Category D.1 in Table 1) can occur among instructors, sections, courses, programs, and schools. Cureton (1971) cites instances over 500 years to illustrate that grading problems are not new. Starch and Elliott (1913) published widely quoted studies of "alarming" variation in grades that different instructors assign to the same papers — in subjectively graded subjects and even in mathematics. Periodic national shifts in grading level are regularly reported — inflation as well as deflation (Willingham & Cole, 1997, p. 305).

Many studies have shown that grades are not comparable among courses (Elliott & Strenta, 1988; Goldman & Hewitt, 1975; Goldman & Slaughter, 1976; Ramist, Lewis, & McCamley, 1990; Strenta & Elliott, 1987; Young, 1990). These analyses have often been based on a comparison of grades after test differences were taken into account. Similar results have been obtained with no reference to test scores by comparing the grades earned by the same students in different courses (Elliott & Strenta, 1988; Goldman & Widawski, 1976). Goldman and Hewitt (1975) proposed a theory of grading variations based on "adaptive level." They suggested that faculty tend to adapt their grading level to the ability level of the students that they usually encounter. Thus, grading tends to be stricter in courses like mathematics and science that often attract stronger students; grading tends to be more lenient in courses like education and sociology that often attract weaker students (Goldman & Hewitt, 1975; Ramist et al., 1990).

A number of studies have demonstrated that noncomparable course grades lower the correlation between test scores and grades within an institution and that adjusting grades to make them more comparable improves the correlation (Elliott & Strenta, 1988; Goldman & Slaughter, 1976; Ramist et al., 1990; Strenta & Elliott, 1987; Willingham, 1985; Young, 1990). The same principle applies to group differences. Using a grade criterion that is more comparable for women and men typically reduces differential prediction (Elliott & Strenta, 1988; Gamache & Novick, 1985; Ramist, Lewis, & McCamley-Jenkins, 1994).

Similarly, variation in grading standards from school to school lowers the correlation between HSA and college grades (Lindquist, 1963; Linn, 1966; Willingham, 1963). Variation in college grading has similar effects on the overall correlation between test scores and performance in college (Astin, 1971; Braun & Szatrowski,

1984). Past work clearly suggests that grading variations are an important source of discrepancies between grade averages and test scores.

The NELS transcript database permits analysis of grading variations across schools and across courses. Variations in grading can be corrected in various ways; for example, by carrying out regression analyses within schools and adjusting the pooled results for restriction in range. Or course-grading differences can be indexed for each student according to the leniency or strictness of grading relative to average test scores in the courses that he or she took. Significant variation in NELS test score scales is unlikely since all students took equated forms of the same test.

Factor 3: Reliability

Measurement error in grades and test scores is separate from systematic variation in grade scales, though these two sources of grade–test score differences may be correlated. While measurement error cannot be indexed by student, it can be taken into account by traditional corrections for unreliability. Test reliabilities are available from NELS. Grade reliabilities can be estimated through appropriate analysis of course-grade records.

Factor 4: Student Characteristics

Both grades and test scores reflect, in varying degree, the knowledge that students acquire in their particular school programs. There is no practical way to gauge differences in the specific knowledge and skills represented in a test and an individual grade average (A.2 in Table 1). It should be possible, however, to find variables that interact with such differences by focusing on student characteristics (B.2 in Table 1) that help to predict differential grade performance; that is, high or low grades in relation to scores on tests covering similar subject matter. Relevant studies can be found in several different areas of research.

The practical interests of college admissions officers have encouraged research on characteristics that might account for students making higher or lower college grades than their admissions tests would indicate. Higher college grades than predicted tend to be associated with positive attitudes about education, confidence in academic work, good study habits, and a willingness to take on demanding courses (Astin, 1971; Willingham, 1965). Poor grades are associated with negative behavior like turning in work late, being tardy, making wisecracks in class, and going to movies frequently (Astin, 1971). In a concurrent analysis of performance in college, several types of behavior were found to be related to differential grade performance: attendance, completing assignments, taking tests on schedule, a study skills scale, taking notes in class, and years enrolled in key high school subjects (Stricker, Rock, & Burton, 1991).

Concurrent analyses of student performance in high school have also identified a number of characteristics related to differential grade performance. Ekstrom, Goertz, and Rock (1988) reported that behavior problems and time on homework showed particularly strong effects. Other significant contributors were school activities, parent aspirations, parent involvement in program planning, and the students' perceived locus of control. Some recent evidence indicates that the portion of homework completed may have a stronger effect on grades earned than does

time on homework (Cooper, Lindsay, Nye, & Greathouse, 1998). One theme is evident in these studies. The students who make good grades in relation to test performance are those who act like serious students. They come to class, participate positively rather than disrupt, and do their homework.

Some have studied extracurricular student activities with a view to broadening the public view of talent, admissions criteria, and useful outcomes of education (Richards, Holland, & Lutz, 1967; Taber & Hackman, 1976; Willingham, 1985). Richards et al. (1967) concluded that nonacademic accomplishments in high school are largely independent of academic achievement, but Werts (1967) challenged that view, citing a correlation of 0.37 between HSA and extracurricular achievement. Hanks and Eckland (1976) reported similar results but sharply distinguished "social" activities (e.g., class leader, editor) from athletic activities. They found that achievement in these two domains correlated, 0.38 and 0.05, respectively, with HSA.

Extracurricular achievement may be a useful variable in identifying possible sources of grade–test score discrepancies if, as Marsh (1992b) argued, such achievement represents "commitment-to-school." Cooper, Valentine, Nye, and Lindsay (1990) suggested that after-school activities have a positive or negative effect on differential grade performance, depending upon whether they contribute to or compete with a student's school work. They found that differential grade performance was positively correlated with extracurricular activities and amount of homework finished, but negatively correlated with watching TV. Data on the effects of employment have been inconsistent (Marsh, 1991).

Studies of "academic self-concept" reflect the common-sense observation that students' attitudes can have a marked effect on their behavior and performance in school (Byrne, 1996; Marsh & Yeung, 1997). In a meta analysis by Hansford and Hattie (1982, Table VIII), self-concept correlated 0.34 with grades and 0.22 with achievement tests. The more specifically the self-concept refers to a particular academic subject, the stronger its relationship to performance in that area (Marsh, 1992a). This domain specificity exacerbates what Byrne (1996, p. 302) called "the most perplexing and illusory" issue in studying the relationship of self-concept to school achievement. Does self-concept cause grades, or do grades cause self-concept? If the former is more accurate, the argument is more convincing that attitudes help to distinguish grades and test scores. Apparently, the effects can work both ways (Byrne, 1996; Marsh & Yeung, 1997).

Eccles (1983) proposed a broader framework of "expectancies, values, and academic behaviors." Her theoretical orientation assumes a very practical network of manifest behaviors and interpersonal relationships — all necessary elements in a student's pursuit of effective personal development through educational achievement. Thus, positive academic attitudes are reflected in aspiration and planning, in recognizing the value of taking certain courses and working hard, in positive relationships with parents and teachers, and in choosing peers who help to define and reinforce academic commitments and habits.

Many studies have shown that educational attainment is related to socioeconomic status and positive relationships with parents (Jencks et al., 1972; Sewell & Shah, 1968). Harris (1995) proposed that peer culture, not family, plays the dominant role

in shaping the behavior of young people. Collins et al. (2000) argued that parents do influence children, but mainly indirectly. With regard to the specific question addressed here, it is not clear whether the family has more effect on grade performance or test performance. Some data suggest the former (Ekstrom et al., 1988).

Factor 5: Teacher Ratings

Several considerations suggest that teacher ratings may help to account for differential grade performance. The most obvious is that teachers know best how students perform on specific instructional goals — outcomes that presumably have a heavy weight on grades earned. It is also possible that teacher judgment can unduly influence grades. The possibility of self-fulfilling halo or bias has long been an active research topic (Brophy, 1983). Another important consideration is the influence of "other factors" in grading. College teachers place value on aspects of complex intellectual performance that are not often reflected in tests emphasizing subject knowledge and skills; for example, curiosity and creativity (Davis, 1965). Due to the multiple purposes of grading in school, teachers report taking into account specific student behaviors in addition to competence in the subject when grades are assigned (Frary, Cross, & Weber, 1993; Robinson & Craver, 1989). Factors worth special attention in this regard include attendance, effort, disruptive behavior, and completing work assignments.

Study Design

Except where otherwise noted, all data were collected in the second followup of NELS in early 1992 when the participants were seniors (Ingels et al., 1994). The general approach in this analysis was to "correct" for the five factors discussed in the previous section by introducing each, in turn, as adjustments or predictors in a multiple regression analysis starting with the concurrent correlation between overall grade average and total NELS Test score. The question was to what extent the *five factors can account for differential grade performance of individuals as well as* groups of students. The database included 10,849 students in regular school programs with test, transcript, and questionnaire data. In order to examine grading differences from school to school, the sample was limited to students attending schools having 10 or more NELS participants with the requisite data. Due to that restriction, the reduced sample included few students who had changed high schools.

This final, reduced sample appeared similar to the sample with full data in most other respects, and an analysis of NELS data in Grade 8 had shown little association between family moves and school performance level (Finn, 1993, p. 66). Nevertheless, due to the sampling limitations, we regarded the analysis as exploratory and did not estimate standard errors of statistics. Results should be interpreted as descriptive of characteristics observed in a database of 8,454 students attending 581 schools throughout the country, who earned some 187,000 credits in 21,000 high school courses.

Variables in the Analysis

NELS students were administered "a series of curriculum-sensitive cognitive tests to measure educational achievement and cognitive growth between the eighth and twelfth grades in four subject areas — reading, mathematics, science, and social studies [history, geography, and civics]" (Ingels et al., 1994, p. 7). Student grades were obtained from the NELS Transcript Component Data File (Ingels et al., 1995) that contained term grades for all courses attempted by each student. NELS slotted each course into a framework of 1,540 academic and vocational courses with different titles. In addition to different types of courses within mathematics, history, science, etc., the framework included placement levels (remedial, honors, etc.) and the year in which the course was taken. NELS also transformed all grading systems to a common scale of 1 to 13 (i.e., A+ to F), which we converted to the more familiar "4-point scale" (A+ = 4.3).

The NELS transcript file contained a great many multiple grades for the same student under the same course title (e.g., repeats, drops, multiple terms). Grades were appropriately weighted and averaged to provide a single grade based on comparable term lengths in each school for all subjects completed. This file provided each student with a measure of total course hours and total course credits on a Carnegie Unit (CU) base, as well as the data necessary for the analysis of course-grading variations. The course grades also served as the basis for HSA(T), reflecting each student's total 4-year transcript (excepting service courses such as physical education and driver training). The correlation (0.62) between this HSA(T) and the unweighted total score on the NELS Test provided the initial baseline relationship between grade performance and test performance for the NELS seniors.

In order to take account of Factor 1 (Subjects Covered), it was also necessary to compute an academic grade average (HSA) based on subject matter generally comparable to that represented on the NELS Test. HSA was the unweighted mean of four subject grade averages that NELS had already computed for each student in four so-called "New Basics" subject areas: English, mathematics, science, and social science (Ingels et al., 1995, p. 56 and Appendix H). Unless otherwise specified, all of the analyses reported here were based upon this set of 490 academic courses. HSA(T) included those 490 plus all work in foreign languages, computer science, many special interest courses of an academic nature, and all vocational courses.

Student Characteristics. In selecting student characteristics, the objective was to identify variables that might help to account for differential grade performance. Based on prior research evidence and data available in the NELS questionnaires and school transcript, the following 26 variables were developed. Numbers 14, 15, and 18 were binary variables. To enhance generalizability and to minimize missing data, most variables were constructed as composites of two to ten types of related information as indicated by the following illustrative material in parentheses. The first three categories are mainly behavior; the last two are mainly contextual.

A. School Skills
   1. Attendance (seldom absent or tardy — from the student and the school record)
   2. Class participation (comes prepared, pays attention, takes notes)
   3. Discpline problems (has trouble with school rules, fights, suspension)
   4. Work completed (turns in work on time, more than required)
   5. Homework hours (per week — at home and school)

B. Initiative
   6. Courses completed (CU — irrespective of grade earned)
   7. Advanced electives (any AP course, took 12th grade Math — though not required)
   8. School activities (number — with added points for awards or offices)
   9. School sports (number — with added points for achievement or leadership)
   10. Outside activities (frequency, awards)

C. Competing Activities
   11. Drugs and gangs (involvement with)
   12. Killing time (with TV, video games, talking, riding around)
   13. Peer sociability (friends like parties, popularity, going steady)
   14. Employment (outside; 20+ hours per week — yes or no)
   15. Child care (in the family; 20+ hours per week — yes or no)
   16. Leisure reading (hours per week, not school related)

D. Family Background
   17. Socioeconomic status (SES) composite (a NELS measure: parents' education, occupation, and income)
   18. Family intact (living with 2 parents or stepparents — from 1990 survey)
   19. Parent relations (discusses school with parents, gets along with parents)
   20. Parent aspiration (want child to continue education)
   21. Stress at home (parent lost job or died; family member uses drugs)

E. Attitudes
   22. Teacher relations (thinks teachers do a good job, solicits their help)
   23. Educational plans (plans college, plans a career that requires college)
   24. Self esteem (pride, optimism)
   25. Locus of control (assumes internal control, believes planning and effort pay off)
   26. Peer studiousness (friends like school, studying, getting good grades)

These measures were developed and retained on the basis of their rationale and substance, not their interrelationships or association with differential grade performance. Thus, the 26 student characteristics originally selected were retained throughout the analysis. Mutually constrained or confounded variables were avoided to minimize collinearity. The mirror image zero-sum issue posed a different problem. In Competing Activities, for example, a student cannot easily score high on killing time, child care, employment, etc., all at once. Each of the variables in that category may affect school achievement, but statistically, they hardly form a coherent construct.

Causality is a special problem. Has involvement in honor societies contributed to a student earning high grades relative to test scores or has such involvement merely resulted from high grades? Thurstone (1947, p. 441) referred to such confounding as "experimental dependence"; that is, misleading correlational results due to artifacts of the situation in which data are gathered. Three levels of potential confounding can be distinguished.

- *Statistical constraint* — automatic and usually identifiable. A failing grade lowers both the student's grade average and the number of credit hours earned.
- *Explicit dependence* — direct but not necessarily identifiable or consistent. Being in a vocational curriculum or an honor society is likely influenced by the grade record.
- *Implicit dependence* — perhaps subtle and unmeasurable, but real. Students come to school, do the work, and continue their education partly on the basis of their grades.

We attempted to avoid the more clearly artifactual and potentially misleading forms of confounding, especially those in the first two categories above. Note, however, that doing well or poorly in school has a reciprocal relationship with a scholastic orientation. Academic motivation is not a momentary state or merely a technical issue. Much of the national effort that goes into encouraging excellence and selecting good students for demanding educational programs is based on that assumption. Finally, the student's interest in making good grades is one of the key phenomena under study here. (See Willingham, Pollack, & Lewis [2000] for further specification of the variables.)

Teacher Ratings. NELS collected a variety of ratings from teachers. We used those ratings that focused on largely observable behavior to minimize the influence of grades. Variables 27 to 31 represent classroom behaviors that might be expected to influence teachers' evaluations of students and are often cited as a legitimate factor in grading:

27. Attendance (seldom absent or tardy),
28. Class behavior (usually attentive, seldom disruptive),
29. Consults teacher (talks with teacher outside class),
30. Educational motivation (works hard, going to college, will not drop out), and
31. Work completed (usually completes homework assignments).

Such ratings were collected in both the sophomore and the senior year. The earlier ratings were used, partly because the data were more reliable and complete (two teachers' ratings on 90% of the sample vs. one teacher's ratings on 70%). Another consideration was less likelihood of earlier ratings being biased by the teachers' knowledge of the students' grade records. Since more than half of the grade record came after the sophomore ratings, they are, to that extent, predictors rather than concurrent evaluations.

Statistical Analysis

Factors 1 through 5 were introduced successively to determine what effect these corrections had on grade prediction and differential prediction and how the factors

operate within groups. Most of these analyses involved familiar applications of regression analysis, which are best described with results. Missing data were not extensive and were handled through pairwise deletion. The analysis of grading variations requires some initial explanation. Linn (1966) described a number of methods that have been used to correct for variation in grading standards. Perhaps the simplest, which we call the "residual method," is to use a correction defined as the difference between the mean actual grade in a course or school and the mean predicted grade. Thus, a School Grading Factor (SGF) and a Course Grading Factor (CGF), are given by

$$SGF_i = \sum_{k=1}^{N_i} [\text{Pred}(HSA_k) - HSA_k]/N_i, \quad (1)$$

$$CGF_j = \sum_{k=1}^{N_j} [\text{Pred}(HSA_k) - CG_{jk}]/N_j. \quad (2)$$

In these equations, HSA_k represents the average grade obtained by the *k*th student in school *i*, and CG_{jk} represents the grade obtained by the *k*th student in course *j*. School-grade residuals for each school and course-grade residuals for each course were determined by subtracting actual grades from predictions [Pred (HSA_k)] based on the regression of HSA on the four NELS tests (original data, total sample). That is, predicted HSA was used as a baseline for determining grading strictness in a particular school or course. The SGF for a given school was assigned to all students in that school. The index for course-grading variations for each student was *M*CGF, the mean CGF for the particular courses that student took. SGF and *M*CGF were expressed on the HSA scale, so that they could be used as predictors or as criterion corrections.

This residual method has two problems. First, it will tend to overfit, because the mean predicted grade for each school or course is identical to the mean corrected grade for that group; that is, the between-group correlation for test mean and grade mean equals 1.0. Therefore, the overall correlation between NELS test score and adjusted HSA is somewhat inflated. Second, the residual method corrects only HSA for scale variations, though scale distortions among schools have also been well documented in other variables important to the analysis. One example is the "big-fish-little-pond" effect whereby students' attitudes about themselves are inversely affected by the average ability of students in their school (Marsh, 1987). School size affects other variables of interest. Larger schools expand course-taking possibilities (Monk & Haler, 1993), but render extracurricular achievements more competitive (Lindsay, 1982). Both the overfitting problem and the evidence of scale distortion in other variables made within-group analysis a more appropriate method for correcting scale differences across schools.

To apply a within-group method, school means are subtracted from the observed scores for all variables (i.e., all grades, scores, ratings, etc.). In this pooled within-school matrix of deviation scores, all variables have a mean of zero — overall and within each school. The pooled within-school correlation matrix based on these deviation scores was corrected for range restriction, using the NELS Composite and the SES Composite as explicit selection variables — the only measures in the

matrix that one might assume to be reasonably comparable across schools. This multivariate correction employs an extension of the Pearson–Lawley method (Gulliksen, 1950, p. 165). A within-school analysis does not entail the overfitting that characterizes the residual method (Willingham et al., 2000), and it removes the "noise" of the scale distortions in other variables as indicated above. The implicit assumption is that the resulting gain in accuracy will outweigh any valid school "signal" that is lost in the process.

Insufficient data prevented using the within-group method to correct course-grading variations. Following Ramist et al. (1994), we used a residual method (a computational variation on Equation 2) that involved adding an adjustment for course-grading strictness to each student's within-school HSA. Due to sparse course data within schools, "course" was always defined as all students taking a subject with the same name, irrespective of school; for example, Algebra I, English II, etc. For the reasons stated, a within-school analysis was our method of choice for correcting school grading variations. The residual method was employed as an additional check, using Equations 1 and 2 in the original across-school data set. The residual method allowed indexing school (SGF) and course (*M*CGF) grading factors as separate variables for individuals. This approach made it possible to study how those factors are related to other variables and, most importantly, to repeat the analysis within individual subgroups.

Results of the Analyses

Accounting for Individual Differences

The correlations in Table 2 show the accumulating effect of correcting Factors 1 through 5 on predicting HSA; that is, the extent to which one can account for observed differences between the grades and test scores of individuals. The eventual multiple correlation of 0.90 was based on the four NELS tests and the five factors as described. This analysis resulted in an apparent increase to 81% of variance accounted from the 38% based on the total NELS Test score alone.

In interpreting these data, it is important to bear in mind several qualifications. The five factors involved different types of corrections: changing the criterion, adding predictors, and adjusting for unreliability. As a result, the total variance that serves as a baseline is not the same at each stage of the analysis and the successive increases in the multiple are not strictly comparable. Furthermore, it is possible to add the factors in a somewhat different order. The end result is the same, but the increases in R associated with each factor can vary. For example, in third position the correction for unreliability increased the multiple by 0.05. Were it Factor 5 at the end, it would increase the multiple only 0.02.

Factor 1 involved two steps to improve the match of subject matter in the grade average and the test score: (a) pairing the subjects in the grade average with those in the test, and (b) optimally weighting the subtests. Weighting the tests first added 0.037; pairing the subjects first added 0.025. Together, they raised the multiple by 0.06.

Factor 2 involved two corrections for variations in grading standards. In a pooled within-school correlation matrix, the range-corrected multiple correlation between

TABLE 2
The Accumulating Effects of Five Factors That Help to Account for Observed Differences Between Grades and Test Scores

R	Status of Grade-Score Relationship
.62	Correlation between total grade average and total NELS score
.68	Plus Factor 1—HSA and test based on similar subject matter
.76	Plus Factor 2—School and course grading variations corrected
.81	Plus Factor 3—HSA and test corrected for unreliability
.86	Plus Factor 4—Student characteristics added
.90	Plus Factor 5—Teacher ratings added

the NELS tests and HSA was 0.75, an increase of 0.07 attributable only to removing the variation in grading from school to school. Correcting for variations in course grading involved adding to each student's grade average a constant, *K*, representing the average strictness of grading in courses taken by that student. The multiple correlation of the NELS tests with this doubly adjusted grade average was 0.76, an additional increase of only 0.01. Using *K* as an independent predictor also had little effect on the multiple *R*. These results are in sharp contrast to the findings of Ramist et al. (1994), who found that a similar index of grading strictness in college courses substantially improved Freshman grade prediction.

One possible explanation is that school and course-grading variations are highly correlated. In that case, the initial removal of school-grading variations in a within-school analysis would have also removed much of the course-grading variations. To examine that possibility, school and course-grading strictness were separately indexed by SGF and *M*CGF, based on Equations 1 and 2 using the original data .The correlation of SGF and *M*CGF was only 0.18. Furthermore, correcting for course-grading variations by using *M*CGF as a criterion correction or as a predictor had little effect on the correlation of test scores with HSA.

Another empirical question concerned the consistency of the pattern of course-grading strictness in high school. Ramist et al. (1994) estimated course-grading strictness within each college. If the pattern of course-grading standards varies noticeably from school to school, the sparseness of NELS data within individual schools would take on added significance. In that case, the lack of sufficient data to represent course by school interactions in grading would limit the effectiveness of a course-grading correction.

To determine whether that was the case, grading variations in the original data set were analyzed by schools, courses, and the interaction of schools and courses. This ANOVA was performed on residual course grades using HSA predictions based on the four NELS tests (see Willingham et al., 2000, for specification of the model). There were 574 schools and 225 courses in the four subject areas on which HSA was based. Table 2 shows the results of the analysis of variation among these school-course cells, weighted to account for unequal *N*s and credits.

Two main findings are evident in Table 3. First, 54% of the variation between cells is attributable to the additive model, which assumes a course effect (represented by *M*CGF) and a school effect (represented by SGF). The course–school interaction accounted for the remaining 46% of the cell variation. Thus, the two

TABLE 3
*Analysis of Course Grading Variations By Schools, Courses, and the Interaction of Schools and Courses**

Source of Variation	Sum of Squares	Proportion of Total
Schools	8115.3	.41
Courses, controlling schools	2751.4	.14
Additive model	10,866.8	.54
Course-school interaction	9118.4	.46
Total between cells	19,985.2	1.00

*Based on 574 schools and 225 courses in four basic academic areas.

grading factors used here actually only identified about half of the systematic variation in grading that is associated with schools and courses. This result indicates that our analysis substantially underestimates the extent to which discrepancies between grades and test scores are due to grading variations.

Second, the ANOVA indicates that our results are misleading regarding the relative effect of course- and school-grading variations. The course main effect, the sole basis of *M*CGF, accounted for 14% of the variation in cell means (i.e., all course-grade residuals). The interaction, representing 46%, is also most reasonably treated as variation in course grading. Thus, differences in the pattern of course grading among schools accounted for three times as much grading variation as did the overall differences from course to course. Both results stemmed from our inability to identify course-grading patterns within individual schools.

Factor 3, Reliability, represents unsystematic error. In meta analyses of validity studies (e.g., Hunter & Schmidt, 1982, p. 88), it is customary to correct only for unreliability of the criterion, but here the objective was to remove the effects of discrepancies between HSA and the test score due to measurement errors in either of the two. Low reliability can be a significant source of assessment errors, especially with a less traditional performance test (Koretz, Stecher, Klein, & McCaffrey, 1994). Measurement error was not a major factor in this analysis, because the reliability of both measures was quite high: 0.97 for HSA, and 0.96 for NELS total score. Both estimates represent the reliability of a composite, using the four subtest reliabilities reported by NELS (Rock, Pollack, & Quinn, 1995) and four split-half grade reliabilities computed for odd–even courses in each of the four subjects represented in HSA (see Willingham et al., 2000).

Previously reported estimates of the reliability of grade averages have often been substantially lower—from the low 0.60s to the low 0.80s (Ramist et al., 1990; Werts, Linn, & Joreskog, 1978). Several explanations are possible. Estimates of the reliability of grade averages have typically involved more limited coursework and samples with a more restricted range. Furthermore, the odd–even estimates here are probably somewhat too high. The test measures performance in the senior year. HSA represents performance through 4 years, but academic achievement can vary from year to year. The odd–even reliability estimate inappropriately ignores annual differences in grade performance instead of treating such differences as error

variance which is implied by Category C.2 of Table 1. Humphreys' (1968, p. 375) report on "a substantial amount of instability of intellectual performance" from year to year suggests that our corrections for grade unreliability are likely to be conservative.

Factor 4, Student Characteristics, involves many achievement dynamics. Columns I and II of Table 4 show correlations of 26 student characteristics with NELS Test and with HSA. As expected, a number of the characteristics were more strongly correlated with grades than with test scores. In two instances, SES and Leisure Reading, the absolute value of the correlation was noticeably lower with HSA than with NELS.

The social advantages implied by a higher SES (Variable 17) would presumably act over a lifetime on the development of general cognitive skill in and out of school. The test is more likely to reflect such general skills than is the school average that focuses on specific learning objectives and behavior in the classroom. Furthermore, the correlation between SES and grade average is likely to be depressed by grading variations (an assumption supported by the corrected correlation in Column III). The correlations with "Leisure Reading" on material unrelated to school (Variable 16) were more surprising. This case may be similar to that of SES. A history of outside reading could raise general cognitive skills, as do the advantages of a high SES, but also take time away from schoolwork. In either sense, leisure reading would be a competing activity with regard to relative performance on grades and tests.

Columns III through VI in Table 4 trace the relationship of each student characteristic with HSA as several corrections and controls are taken into account. These four columns are based on within-school analyses, corrected for range restriction. As expected, correcting grading errors in Column III often increased somewhat the correlation with HSA. Column IV shows the partial correlation of each student characteristic with HSA when grading variations were corrected and test scores were controlled. This partial correlation gives the best indication of the extent to which each characteristic is related to differential grade performance.

The β weights in Column V indicate which student characteristics made an independent contribution in accounting for grade performance. The interesting point at this stage of the analysis was what types of variables remained in the picture. Family Background and Student Attitudes (Variables 17 to 26) largely dropped out. It was mostly the behavioral variables — especially those directly involved with school — that made an independent contribution in accounting for grades earned. The variables in Categories A through C all involve behaviors that represent different aspects of being engaged in school: employing appropriate school skills, demonstrating initiative, and avoiding competing activities. Nine of these 16 behavioral measures (marked #) showed a partial correlation with HSA of $\pm$ 0.10 or larger when grading variations and test scores were held constant. In subsequent analyses, a composite—termed *Scholastic Engagement* — was based on these nine measures, and weighted in proportion to their partials.

Factor 5, Teacher Ratings, showed strong relationships with grade performance and with differential grade performance. Four of the five ratings correlated from 0.37 to 0.63 with HSA. A single 5-point rating on whether the student completed

TABLE 4

The Relationship of Student Characteristics and Teacher Ratings to the NELS Test (NELS-T) and to High School Average (HSA) with Progressive Controls Applied

Variable	*R* with NELS-T	*R* with HSA	*R* with HSA	β weight: Predict HSA, Reliability Corrected	
			Grading Control	Test & Grading Control	First 26 Variables	All 31 Variables
	I	II	III	IV	V	VI
A. School Skills						
1. Attendance	.19	.33	.39	.31#	.11*	.07*
2. Class participation	.09	.21	.23	.23#	.01	.00
3. Discipline problems	−.22	−.31	−.35	−.23#	−.03*	.00
4. Work completed	−.04	.18	.18	.30#	.11*	.09*
5. Homework hours	.22	.21	.21	.06	−.04*	−.04*
B. Initiative						
6. Courses completed	.28	.32	.45	.24#	.07*	.04*
7. Advanced electives	.59	.58	.65	.34#	.16*	.12*
8. School activities	.19	.28	.32	.18#	.03*	.03*
9. School sports	.04	.06	.06	.03	−.00	−.01
10. Outside activities	.15	.14	.18	.06	−.01	−.01
C. Competing Activities						
11. Drugs/gangs	−.07	−.20	−.24	−.21#	−.02	−.00
12. Killing time	−.10	−.16	−.16	−.14#	−.03*	−.03*
13. Peer sociability	−.09	−.09	−.11	−.05	−.02	−.01
14. Employment	−.12	−.13	−.13	−.08	−.01	−.01
15. Child care	−.10	−.07	−.08	−.01	−.00	−.00
16. Leisure reading	.20	.06	.07	−.08	−.06*	−.04*
D. Family Background						
17. SES composite	.48	.35	.43	.11	.03*	.02
18. Family intact	.11	.13	.15	.08	.02	.01
19. Parent relations	.07	.17	.20	.17	.02	.01
20. Parent aspiration	.34	.33	.39	.15	−.00	−.01
21. Stress at home	−.07	−.12	−.13	−.10	−.00	−.00
E. Student Attitudes						
22. Teacher relations	.21	.23	.19	.12	.01	.01
23. Educational plans	.33	.35	.41	.21	.04*	.03*
24. Self esteem	.28	.29	.35	.17	.00	.00
25. Locus of control	.24	.25	.27	.13	.02	.02
26. Peer studiousness	.25	.29	.30	.18	−.01	−.01
F. Teacher Ratings						
27. Attendance	.22	.37	.43	.31		.01
28. Class behavior	.35	.51	.55	.41		.04*
29. Consults teacher	.11	.15	.18	.10		−.01
30. Educational motivation	.45	.63	.68	.50		.12*
31. Work completed	.33	.61	.65	.56		.20*

#Variables used to define Scholastic Engagement

$N = 8454$; *$\beta \geq .03$.

assignments had a partial correlation of 0.56 with HSA, with grading variations and test scores controlled — the highest partial for any variable that was examined. Three of the five teacher ratings had significant weights in the overall regression analysis (see Column VI of Table 4). Teacher Ratings raised the multiple correlation 0.04 (0.86 to 0.90), even after all other factors were taken into account.

In Variables 27 to 31, teachers largely described what students did in school, and these ratings overlapped with some of the student self-ratings. Columns V and VI of Table 4 indicate that the Teacher Ratings accounted for some, but not all, of the predictive variance associated with the student characteristics. The larger weights in the analysis that was based only on the 26 student characteristics dropped somewhat when the Teacher Ratings were added, but not nearly to zero. Thus, information from both teacher and student helped to account for differences in grade and test performance. For use in subsequent analyses, a Teacher Rating Composite (TRC) was based on the best-weighted average of the five ratings for predicting HSA.

Confounding between grades and predictors is a legitimate concern. We examined each of the 31 predictors in Table 4 for possibly reflecting grades rather than accounting for grades. Three variables seemed most likely to have some confounding with HSA: 7. Advanced electives that attract students with good grades, 23. Educational plans that are likely to be optimistic if the grade record is good, and 30. Educational motivation that teachers may rate high simply because the grades are higher. When all three of these variables were removed from the final regression analysis shown in Table 4, the multiple correlation was reduced by only 0.007. Since other variables largely accounted for the predictive variance contributed by these three measures that seem to be the most suspect, the role of confounding does not appear to be large.

Differential Prediction

While grade prediction indicates the level of grade–test score discrepancies for individuals, differential prediction gives similar information for groups. Table 5 shows the extent to which the HSA of eight groups was over- or underpredicted on the basis of NELS tests and the accumulating effects of adjusting for grading variations, student characteristics, and teacher ratings. Predictions were based on the regression line for the total sample.

Among the eight groups, initial differential predictions based on the NELS tests alone ranged from +0.13 to −0.14; the average was 0.09, disregarding sign. Results converged toward zero as Grading Variations, Student Characteristics, and Teacher Ratings were taken into account. With these corrections, differential prediction diminished in all cases where there was originally any differential prediction. After adjustment for all three factors, the absolute level of differential prediction averaged about 0.02 of a letter grade.

The three corrections affected the groups somewhat differently. Grading variations had a small effect on differential prediction for most groups but a fairly large effect in reducing overprediction for African-American students. These somewhat surprising results are inconsistent with an assumption that African-American and Hispanic students are likely to benefit from easy grading due to being overrepre-

TABLE 5
Differential Prediction for Four Subgroups and Four School Programs—By Amount of Predictive Information

	Predictive Information*			
	NELS Test	Plus Grade Variations Controlled	Plus 26 Student Variables	Plus Teacher Judgments
Group				
Women	+.12	+.11	+.06	+.03
African American	−.10	−.01	−.03	−.02
Asian American	+.13	+.14	+.06	+.04
Hispanic	−.05	−.03	−.03	−.02
School Program				
Rigorous Academic	+.12	+.13	+.06	+.04
Academic	.00	.00	−.01	−.01
Academic-Vocational	+.03	−.00	+.03	+.02
Vocational	−.14	−.13	−.02	+.01

*Entries are mean actual grade minus mean predicted grade. Predictions take into account predictors used in previous columns. $N = 7571$ in Col. 1–3; 6853 in Col. 4.

sented in poor schools. In fact, African-American students were slightly overrepresented in schools that graded more strictly.

Taking student characteristics into account had the largest effects on differential prediction. Underprediction was reduced (by −0.05 to −0.07) for three groups: women, Asian Americans, and students in Rigorous Academic programs. These groups tended to achieve higher grades than one might expect on the basis of the NELS Test. Correcting for student variables had a more substantial effect (+0.11) in accounting for overprediction of the grades of Vocational students. These effects were notably associated with different levels of Scholastic Engagement. For example, the standard mean gender difference on Engagement was 0.48. Teacher Ratings tended to reduce differential prediction slightly for all groups.

A Condensed Analysis of Major Factors

The results to this point suggested that each of the major factors contributing to differences between grades and test scores might be represented with little loss by a single variable or composite. A condensed analysis of the major variables can be helpful in clarifying relationships among the several main factors and in testing the generality of the results within subgroups of students. Since there are no school differences in a within-school data matrix, it was necessary to use an across-school analysis so that school and course grading variations could be indexed as separate predictors. This step prompts the question as to how results compare in a within-school and an across-school analysis.

In an analysis based on 37 variables (31 student and teacher variables, four tests, and two grading factors), the across-school (residual) method gave quite similar results to the previous within-school analysis. As expected, it was slightly less successful. The multiple R was 0.88 in the across-school analysis, compared to 0.90

TABLE 6
*A Condensed Analysis: Intercorrelations and β Weights for HSA Regressed on Four Major Factors**

	NELS-C	SGF	Engage	TRC
School Grading	.09			
Engagement	.41	−.05		
Teacher Rating	.49	−.08	.52	
HSA	.71	−.29	.57	.69
Multiple *R*	β Weight			
.876	.50	−.30	.18	.33

*NELS-C (NELS composite) and HSA were corrected for unreliability.

in the original analysis. Evidently the more thorough scale corrections in the within-school method enhanced *R* more than did overfitting in the residual method. The pattern of β weights was highly similar in the two methods. Next, a condensed regression analysis was based on these four composite variables:

- NEL-C, the best weighted composite of the four NELS Test predictors of HSA, incorporates Factor 1 in the model.
- SGF largely represents Factor 2 since the effects of course grading variations were not consequential in either of the earlier analyses.
- Scholastic Engagement, the nine behavioral characteristics most associated with differential grade performance, largely accounts for Factor 4.
- TRC, the best-weighted rating predictors of HSA, incorporates Factor 5 in the condensed model.

Table 6 shows correlations among these variables and HSA. Since the correlations are corrected for unreliability of the test and HSA, Factor 3 is incorporated in addition to the four factors represented by the four composite variables. The noteworthy multiple *R* of 0.876 is within rounding error of the *R* based on all 37 predictors in the comparable across-school analysis. Each of the three non-test variables correlated substantially higher with HSA than with NELS-C, and each had a consequential β weight in the multiple regression.

Gender, Ethnicity, and School Program

Table 7 shows multiple regression results in four-variable analyses for two gender, four ethnic, and four program groups. Results were highly similar from group to group. The multiple *R*s were all at nearly the same level. As in the total sample, little information was lost in the subgroup analyses by going from the 37-variable to the 4-variable analysis. With little exception, the standard regression weights were also quite similar across groups. The NELS Test and the Teacher Rating, in particular, carried almost exactly the same weight in all gender and ethnic groups. Indeed, the only differences of any note were a slightly larger weight for Engagement and a somewhat smaller weight for School Grading in the case of Asian-American students.

TABLE 7
*Analyses of HSA Regressed on Four Major Factors—By Gender, Ethnic Group, and School Program**

	Beta weight, Predicting HSA				Multiple *R*	
	NELS Test	School Grading	Engage-ment	Teacher Rating	4 Major Factors	All 37 Variables
Gender						
Male	.49	−.30	.16	.34	.86	.88
Female	.53	−.31	.17	.30	.89	.89
Ethnicity						
African Amer.	.50	−.31	.16	.34	.85	.87
Asian Amer.	.50	−.23	.24	.31	.87	.88
Hispanic	.48	−.34	.20	.31	.83	.84
White	.49	−.30	.17	.34	.88	.89
Program						
Rigorous Acad.	.57	−.32	.15	.31	.88	.89
Academic	.52	−.31	.18	.31	.87	.88
Acad.-Vocat.	.50	−.33	.16	.34	.83	.86
Vocational	.35	−.32	.11	.44	.73	.77

*Corrected for range restriction in all variables and unreliability in HSA and NELS Test. (Total $N = 8454$ except for teacher rating where $N = 7619$)

Regression analyses also showed such similarities among the four school programs, though changes were evident in the less academic curricula, especially the Vocational program. HSA tended to become less predictable and the pattern of weights shifted even though HSA was based only on academic courses. NELS Test became a less important predictor, Teacher Rating more important. That pattern may be due to emphasis on different competences in the academic courses taken by Vocational students. Social promotion may also play a role. The practice of passing less academically inclined students partly on the basis of effort (Public Agenda, 2000) could produce grades that correlate somewhat less with tested knowledge and skill.

Table 8 shows group means for these measures on a standard scale. The school-based grading factor varied little from group to group. HSA and the other three composites reflect individual differences. Those measures give somewhat different views of student performance in school. Score levels differed substantially from program to program in a rather consistent manner: Students in a given program had similar mean scores on each of the four measures, but the means were progressively higher in the more academic areas.

Means for the gender and ethnic groups tell a similar story with some exceptions. Mean scores were almost as divergent in the four ethnic groups as in the four school programs. Each group tended to show a characteristic pattern, above or below average to some degree, though Hispanic and African-American students scored somewhat lower on NELS Test and HSA compared to Engagement and Teacher Rating. Since these groups may be more likely to cluster in particular schools, this latter finding could be due partly to lack of comparability in teacher

TABLE 8
*Standard Mean Scores on Major Factors Related to Differences Between Grades and Test Scores—By Gender, Ethic Group, and School Program**

	N	HSA	NELS Test	School Grading	Engage-ment	Teacher Rating
Gender						
Male	4154	48.6	50.2	50.3	47.6	48.1
Female	4300	51.4	49.8	49.7	52.3	51.8
Ethnicity						
African Amer.	591	43.8	42.7	52.7	49.7	47.0
Asian Amer.	524	54.4	54.1	50.5	53.0	53.8
Hispanic	804	46.0	45.0	50.5	47.7	47.6
White	6471	50.8	51.0	49.7	50.1	50.3
Program						
Rigorous Acad.	1892	55.1	55.2	50.7	54.4	54.4
Academic	4650	50.7	51.0	49.9	50.6	50.6
Acad.-Vocat.	627	46.6	44.4	48.7	47.7	47.1
Vocational	402	41.5	40.2	50.0	42.2	41.9

*Each measure was scaled to a mean of 50 and a SD of 10 for the full sample. Total $N = 8454$ except for Teacher Rating where $N = 7619$.

ratings and student responses from school to school. A more notable finding is the similar mean performance on test score and grade average within each ethnic group. Results here show the familiar tendency of women and men to score somewhat better on grades and tests, respectively (0.16 *SD* in each case). That gender difference may be largely due to women being more involved in school. Mean scores favored women on all nine components of Scholastic Engagement.

Discussion

In evaluating these results, recall two limitations of the analysis. The data concern high school performance. At lower grade levels or in advanced education, the relationship between grades and test scores will engage additional issues not considered here. Also, our focus was on why students score differently on grades and tests, not on the specific content of the measures or what each ought to represent from a social or educational perspective. Average differential prediction of 0.02 letter grades suggests little room for improvement in accounting for differences in these commonly used grades and test scores. But a correlation of 0.90 leaves 19% of the grade variance of individuals unaccounted for. Why? Other likely sources suggest that the differences remaining are neither obscure nor unreasonable.

First, grades are based on a syllabus that varies to some degree across schools, classrooms, and individuals, but tests are designed to avoid content that is unique to particular learners or learning situations. Thus the test content is constant, but the substance of the grading standard necessarily varies from student to student. There is no way to adequately "correct for" that fundamental difference between grades and tests. Second, teachers' grades and external tests naturally focus on somewhat

different constructs. Grades reflect a broader range of knowledge and skills than can be represented in a limited test with restricted modes of assessment.

Furthermore, students do not perform the same in school from one year to the next. We apparently underestimated the effects of such unreliability in grades. There were other technical shortcomings. The accuracy of questionnaire data is always open to question, especially data from graduating seniors. The Teacher Ratings were unreliable and came early in the students' high school program. Finally, the total effect of grading variation was underestimated.

The Problematic Variation in School Grading

Our finding of significant grading variations among schools is consistent with the research cited earlier. Grading probably also varies from one instructor to another, but with this database one cannot tell. We do know that different patterns of course-grading variations from school to school was a major source of uncorrected grade–test score differences. That result was unexpected but instructive. Unlike the findings of previous work on college grades (Elliott & Strenta, 1988; Ramist et al., 1994; Young, 1990), we found little grading variation among high school courses overall. The different result is apparently due to the necessity to pool course-grade data across schools in this analysis, while in earlier studies it was possible to analyze course-grade variations within each institution. In our data, ANOVA results indicated that most of the substantial course-grade variations were hidden in differences among course-grading patterns from school to school.

These findings need confirmation, but in any data similarly limited, the quality of a course-grading adjustment so derived is likely to be similarly degraded. Information about course grading in individual schools is limited in most high-stakes admissions decisions. Our data indicate no sound basis for knowing whether a course with a particular title will be strictly or leniently graded in a given secondary school. This ambiguity makes fair interpretation of course grades and grade averages doubly problematic, especially if grades are used alone. Apparently, even if a given school is known to grade strictly or leniently overall, one can not confidently say that the school's grading effect applies to a given student or to a particular course grade.

Scholastic Engagement as an Organizing Principle

The idea that learning does not depend upon cognition alone is hardly new. A century ago, Dewey (1900) rejected the image of the learner as an empty vessel to be filled through instruction. By mid-century others had recognized the critical roles of expectancies, values, and situational determinants (McClelland, Atkinson, Clark, & Lowell, 1953). These ideas framed a broadened domain of research and theory on achievement and its motivation (Pintrich & Schunk, 1996; Weiner, 1992). Snow (1989), in particular, articulated the conative aspects of learning — notably such factors as interest, volition, and self-regulation.

Several writers have referred to "involvement" or "engagement" in different ways, to characterize various student attitudes and behavior in school that foster academic achievement and personal development (Eccles, Wigfield, & Schiefele, 1998; Finn, 1993; Lamborn, Mounts, Steinberg, & Dornbusch, 1991). Our

analysis started with a number of student characteristics that had previously shown promise as individual measures. Those characteristics that reflected overt behavior made most of the independent contributions to differential grade prediction and provided a rational basis for effective "Scholastic Engagement." Three components described a recognizable pattern:

- The engaged student employs appropriate School Skills. Engagement means coming to school, participating in class, refraining from misbehavior, and doing the work assigned.
- The engaged student takes Initiative in school. Engagement means taking a full and demanding program of coursework and participating in other scholastic activities. Adelman (1999) provides recent corroborative evidence on the value of a robust transcript.
- The engaged student avoids unnecessary Competing Activities. Engagement means abstaining from pursuits that take undue time and commitment away from schoolwork; for example, killing time and involvement with drugs and gangs.

The behavioral focus is a distinct benefit. To be sure, the behavior of students is constrained by background and conditioned by peers. Nevertheless, it is reasonable to assume, as the data suggest, that behavior most directly affects achievement. Also, behavior is more readily described and observed and can, in principle, be modified to the benefit of teachers and learners. If so, the idea of behavioral engagement in school can be a useful tool in studying and improving the educational process. But if students do well because they are engaged, is it not just as likely that they are engaged because they do well? Is the engagement argument circular? It is to a degree, of course. But, from an educational perspective, the arguable challenge is to influence student behavior in ways that create positive self-fulfilling prophecies of achievement and development.

Group Performance: Different Levels, Similar Dynamics

Fairness in high-stakes tests is often cast in terms of differential prediction and differential validity among groups (Linn, 1982). On average, groups performed at somewhat different levels but tended to have similar scores on grades and tests. Results of differential prediction suggested that even the differences observed could be largely accounted for when important non-test variables were taken into account. It is useful to think of differential validity as pertaining to possible differences in the dynamics of achievement. Are the major variables that influence achievement similarly related and similarly constituted from group to group?

Indeed, among the six gender–ethnic subgroups, interrelations among the major variables were very similar, prediction of HSA gave quite similar multiple correlations, and regression weights for the major variables had an almost identical rank order for each group. The major variables were also quite similarly constituted. For example, among the five components of the TRC, Does Homework and Educational Motivation had the first and second largest regression weights, respectively, for predicting HSA in all six gender and ethnic groups. Similarly, among the nine characteristics defining Scholastic Engagement, the same three measures (Ad-

vanced Electives, Attendance, and Work Completed) accounted for the top three partial correlations with HSA with little exception from group to group.

Such results suggest that the dynamics of achievement are quite similar across gender and ethnic groups. The four school programs typically gave similar results. What is the implication of similar dynamics? Consistency in the way that variables operate from group to group indicates little differential validity — an important mark of fair assessment. Groups do differ in interests, background, and culture, but these results give little indication that such differences impart a different meaning to grades, test scores, and other major variables that influence school achievement.

Validity and Fairness

What do the findings suggest regarding the merits of grades and tests for high-stakes decisions? Grades and tests should be mutually validating if they are differently derived indicators of much the same performance. Anything less than a strong correlation between test results and grade results is usually taken to be evidence of invalidity and unfairness (in test scores, seldom in grades). That interpretation seems inconsistent with the results reported here. Taking into account reasonable sources of difference, we found grades and test scores to be strongly related — for individuals as well as groups. Furthermore, the several corrective factors appeared to work in a largely similar fashion from group to group. Given a grade average and a test score based on generally similar subject matter, discrepancies between the two appear to have less to do with mysterious sources of invalidity or defects in the test than with errors in the grades and incomplete information about the students and their behavior in school.

But do the findings suggest that, except for a few known differences that can possibly be taken into account, grades and tests based on a corresponding domain are likely to be comparably valid and fair? In a sense that is true, but that view of the topic leaves a lot unsaid. The corrections themselves bear directly on validity and fairness. Also, the statistical criteria warrant a second look. Finding it possible to account for much of the apparent discrepancy between grade and test performance should caution against excessive reliance on the statistical indices that we seek to explain and improve.

Researchers argue whether predictive validity is 0.50 or 0.55, whether differential prediction is 0.10 or 0.05 of a letter grade even though technical artifacts influence the indices. The artifacts include range restriction and other aspects of sampling (Lewis & Willingham, 1995; Linn, 1983), unreliability of predictors and criteria (Humphreys, 1968; Linn & Werts, 1971), relevance and comparability of the criterion (Elliott & Strenta, 1988; Ramist et al., 1994; Willingham, 1985), whether other appropriate predictors are included in the analysis (Linn & Werts, 1971), and institutional differences and changes over time (Willingham & Lewis, 1990). Typically, such artifacts lower validity coefficients and increase differential prediction. They largely concern the appropriateness of the grade criterion and the nature of the sample, which are not features of the test.

In principle, any of the factors in this analysis or in Figure 1 could be considered an artifactual source of grade–test score differences in a particular assessment. For example, differences between grades and test scores because teachers adjust stu-

dents' grades for attendance and class behavior would be artifactual in assessing knowledge and skills required for high school graduation. But errors due to variation in grading standards from group to group or unreliable grading within a group would be artifactual differences in most assessment situations. The results here demonstrate the hazards in uncritical interpretation of traditional indices of validity and fairness, because it is clearly possible to substantially increase grade–test correlations and decrease differential predicition with corrections that may have little if any actual connection with the quality of the test being evaluated. These common statistics will give a misleading and unduly conservative picture of validity and fairness if artifacts are not taken into account. Furthermore, the statistical indices are an insufficient standard for judging validity and fairness.

A broader view of the topic must include the long-term consequences of using a test, not just the particular high-stakes decision at issue. These include the backward effects on instruction and learning and the forward effects on the eventual social outcomes of education (Frederiksen, 1984; Resnick & Resnick, 1992; Shepard, 1992b). A correlational model has inherent limitations. Our analysis tells little about the relevance or sufficiency of the knowledge and skills that are normally represented by grades and tests. Being able to correct a major part of observed differences between grades and test scores does not mean that they are the same. To the contrary, corrections mean dissimilarities and suggest that, in part, the merits of grades and tests rest on different strengths.

Differential Strengths

Of the five factors originally included in the analysis, three represent characteristic differences between grades and test scores: Factor 2, grading variations; Factor 4, scholastic engagement; and Factor 5, the teacher's judgment of the student's performance. Each of these can be seen as a component of grades that is less represented in tests, if at all. Table 9 focuses on possible substantive strengths of grades and tests that may be implied by Factors 2, 4, and 5. Two other factors—Factor 1, Subject Match and 3, Reliability — are not considered here because neither consistently represents a characteristic difference between grades and test scores.

Each of the three factors in Table 9 is associated with a component of grades that is pertinent to an assessment objective. They concern fairness, fostering development of critical skills, and motivating teaching and learning. Table 9 also takes account of the context in which assessment is carried out. A grade represents each teacher's judgment as to how well a student has fulfilled the implicit local contract between teacher and student. For example, the understanding may be, "If you master the particular knowledge and skills pertinent to the objectives of my course, you will probably get at least a B — maybe higher if you do all the assignments and contribute to the class, but maybe lower if you neglect your homework and disrupt class."

A test, on the other hand, provides an external standard that is intended to compare performance across educational units. For that reason, the test is designed to include important knowledge and skills that are common to relevant coursework. Table 9 connects substantive issues (the goals and content of education) with

TABLE 9
A Schema for Considering Possible Strengths of Grades and Test Scores in High Stakes Decisions

Differentiating component in grades	Pertinent assessment objective	Differential Characteristics: Grades (as performance on the local contract)	Differential Characteristics: Test Scores (as performance on the external standard)
Grading Variations (Factor 2)	To assess fairly	Local standards	Common yardstick
Scholastic Engagement (Factor 4)	To evaluate critical skills	Conative skills	Cognitive skills
Teacher Judgment (Factor 5)	To motivate teaching and learning	Assessment based on each student's learning and behavior	Assessment based on designated knowledge and skills

performance issues (the statistics of individual differences). The schema helps to distinguish some important strengths in assessment, though the distinctions are not exact or exhaustive, nor do they apply equally to all types of high-stakes decisions in education.

One objective in high-stakes assessment is to insure that all students are evaluated on the same scale (relevant to Factor 2). In the case of grades, the same scale usually means the local standard. Sometimes that may be viewed as sufficient for fair assessment. Passing standards that vary from one educational unit to another are not necessarily dysfunctional. Programs with students who are academically weak or unusually talented may not be well served by the same grade scale. On the other hand, some high-stakes decisions call for assessment that is comparable across educational units — either to be fair to all students or, for educational purposes, to base decisions on the same standards of competence. Having a common yardstick is a principal reason why tests are used. To be sure, test score scales are not always dependable. Some K-12 tests have been known to deliver suspicious scores (Cannell, 1988). But in most high-stakes situations, consistent scale meaning is likely to be a major strength of tests because they can compensate for a grade average that may be either inflated or overly strict.

Another objective of high-stakes assessment is to evaluate critical outcomes of education (relevant to Factor 4). Even if grades and tests are based on the same subject matter, there are clearly important distinctions in the content. Table 9 contrasts conative and cognitive skills; this content difference is a critical distinguishing mark of construct relevance for a high-stakes measure. Grades especially reflect engagement, a strength particularly associated with conative skills. Educators recognize that conative skills like volition, habits of inquiry, effort, and self-regulation are, in themselves, important goals of schooling in a free and

effective society (Krathwold, Bloom, & Masia, 1964; Snow, 1989). The focus on cognitive skills is a strength of tests. Currently, the popular sentiment is to hold students to a test-based, largely cognitive graduation requirement (Baker & Linn, 1997). Recent research has added to our understanding of a broader range of cognitive skills that are valuable in academic work and adult life (Shepard, 1992a). Much effort has also gone into improved design and delivery in the assessment of such skills (Frederiksen et al., 1990; Linn, 2000; Mislevy et al., 1999).

Finally, a time-honored objective of educational assessment is to motivate achievement by providing feedback (relevant to Factor 5). "Motivate" implies the need to shape learning as well as encourage effort. Grades and tests perform this function differently. Teachers' grades provide an immediate reward or punishment to students for good or poor work on the specific material assigned in each classroom. External tests are more concerned with how students are doing at the end of the year on knowledge and skills that are most relevant and common to the curriculum — in the system or the state, depending on the test. Shepard (2000) referred to this content distinction as the formative and summative roles of classroom assessment and external tests, respectively. It is no surprise that teacher ratings are more highly correlated with grades than with test scores. Teachers often report passing students' if they have tried hard (Public Agenda, 2000). Parents and teachers hope that grades will motivate students to do their lessons, week by week. Administrators and politicians hope that annual results on common tests will motivate teachers and schools by identifying the most important outcomes and encouraging accountability.

Frequent grades provide more regular feedback on a broader range of skills than annual tests. The added breadth that teacher marks add to assessment may well come at the expense of the consistent meaning of the grade. But tests can also be off the mark. Heubert and Hauser (1999) describe the public policy dilemmas and the legal issues that arise if an external test does not adequately match the school's curriculum or is otherwise considered unfair to its students. Thus the direct content relevance of grades and the focused comparability of test scores are important complementary strengths of the two measures.

The TRC correlated 0.69 with HSA. Is all of that strong correlation good news? Considering that the average of only 1.6 teacher ratings per student came from the sophomore year, this substantial relationship is not likely to be explained simply as confounding. Could the 0.69 partly reflect a tendency for grades to be influenced by teachers' opinions of who deserves good grades? A high correlation between grades and ratings of behavior further illustrates the need for more research on the validity and fairness of grades for high-stakes decisions. More balanced concern would also be desirable regarding the fair and legal use of grades as compared to the obligations of test use (Office of Civil Rights, 1999).

We have noted these differential strengths of grades and tests: Grades can represent broader content and reflect unique accomplishments, but tests can more easily emphasize the most important content. Tests can more readily assess cognitive skills, but grades can more readily assess motivational components of achievement. Grades can reflect progress on what each student is studying, but tests can reflect progress on more significant long-term educational objectives. Test scores

are more comparable from one school to another, but grades scales are more readily accommodated to local situations and programs. The value of these distinguishing strengths will naturally depend upon the specific situation and the purpose of an assessment, but the strengths of grades and tests clearly complement one another.

For that reason, it will often prove desirable to use grades and tests together in making consequential decisions about individual students. But this analysis has been based on grades and tests that are familiar and commonly used. There is no reason to assume that either grades or tests are now as good as they could be or need to be. Measurement specialists and educators might do well to worry less about the statistical indices of the grade–test relationship and more about what is assessed and how that can improve teaching and learning.

Summary

Both grades and test scores play an important role in high-stakes educational decisions. Tests are often used because of uncertainty about the meaning of grades, yet grades are used to evaluate the validity and fairness of tests. Grades and tests provide this mutual support because it is commonly assumed that they do or should measure much the same thing. Yet the two measures often yield somewhat different results. This study is an attempt to account for differences between grades and test scores and thereby improve understanding of their merits as high-stakes measures.

A framework of likely differences between grades and test scores was proposed. To evaluate how well the framework accounted for differences between the NELS Test scores and HSA, a five-factor approximation was based on these steps: (a) focus the grade average and the test score on similar subjects, (b) correct both for unreliability, (c) correct for grading variations, (d) add 26 relevant student characteristics, and (e) add five teacher ratings of student academic behavior. These adjustments improved concurrent prediction of HSA to 0.90 compared to 0.62 based on the test alone. Average differential prediction for eight subgroups was reduced to 0.02 of a letter grade. The improvement in grade prediction was achieved almost equally well by representing Factors 3, 4, and 5 with only three major variables in place of the 33 originally used.

Grading variation among schools was one major source of discrepancy between observed school grades and grades predicted from test scores. Course-grading variations were also found to be substantial, but largely uncorrectable because the data were insufficient to adjust for variations in course-grading patterns from school to school. Difficulty in identifying such patterns in practice may make fair interpretation of grade records even more problematic.

Another major factor in grade–test score discrepancy was Scholastic Engagement, a composite variable related to both measures, but especially to grades. Engagement was defined by three types of observable behavior: employing appropriate school skills, showing initiative in school, and avoiding competing activities. Because of its intrinsic relevance and apparently critical role, Scholastic Engagement shows promise as an organizing principle in studying and improving school achievement. Teacher Ratings were another major factor in accounting for grade–test score differences, partly because teachers often take student behavior directly

into account in assigning grades. Student attitudes and family variables appear to play an indirect role through their influence on student behavior.

Gender, ethnic, and program groups differed in average achievement level, though the pattern of group performance was generally much the same on grades and tests. While achievement level varied, achievement dynamics were mostly similar among groups. That is, the variables accounting for individual differences in grades and test scores were similarly constituted and functioned in a similar manner from one group to another, especially among the six gender and ethnic groups. Typically, tests played the principal role in accounting for grades, though less so in the case of vocational students.

The fact that individual and group differences in grades and test scores can be largely accounted for by a few sensible factors that work in very similar fashion for different groups of examinees speaks to the intrinsic validity and fairness of both grades and tests. This analysis and other work cited show that common statistical indicators will give an unduly conservative picture of validity and fairness unless artifacts are taken into account — which is usually not done and often not possible in practice. That circumstance illustrates the important limitations of traditional indices of validity and fairness and stresses the need to understand what skills are represented in a high-stakes measure. Finally, the corrective factors themselves show that grades and tests have different strengths with strong implications regarding the validity and fairness of each measure.

Grade performance and test performance appear to be systematically distinguished by three added components in grades that are less salient or missing altogether in test scores: the variations in grading standards, the conative skills in scholastic engagement, and the influence of teacher judgment in representing local educational objectives. These three components tend to be associated, respectively, with overlapping but somewhat different objectives in high-stakes assessment: using measures that are comparable for all students, representing the most critical skills, and motivating the educational process.

Grades motivate students through the local contract with the teacher; tests motivate teachers and schools through the external standards thereby imposed. Due to their distinguishing characteristics, grades and tests have different strengths that tend to be complementary. Common advice that the two measures should be used together where possible is well founded. Emphasis on research and other efforts that might enhance their complementary strengths would be well placed.

Note

The authors are grateful for the useful suggestions of reviewers Henry Braun, Brent Bridgeman, Nancy Burton, Robert Linn, and Lawrence Stricker and for the editorial assistance of Linda Johnson. Appreciation is also expressed to National Center for Educational Statistics for providing the data for the analysis and to Educational Testing Service for supporting the study.

References

Adelman, C. (1999). *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment.* Washington, DC: U. S. Department of Education, Office of Educational Research and Improvement.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Astin, A. W. (1971). *Predicting academic performance in college.* New York: The Free Press.

Baker, E. L., & Linn, R. L. (1997). *Emerging educational standards of performance in the United States* (CSE Technical Rep. 437). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Beatty, A., Greenwood, M. R. C., & Linn, R. L. (Eds.). (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions.* Washington, DC: National Academy Press.

Bennett, R. E., & Ward, E. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Braun, H. I., & Szatrowski, T. H. (1984). The scale-linkage algorithm: Construction of a universal criterion scale for families of institutions. *Journal of Educational Statistics, 9*(4), 311–330.

Brophy, J. E. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology, 75*(5), 631–661.

Byrne, B. M. (1996) Academic self-concept: Its structure, measurement, and relation to academic achievement. In B. A. Bracken (Ed.), *Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 287–316). New York: John Wiley & Sons, Inc.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practices, 7*(2), 5–9.

Collins, W. A., Maccoby, E. E., Steinberg, L., Hetherington, E. M., & Bornstein, M. H. (2000). Contemporary research on parenting: The case for nature and nurture. *American Psychologist, 55*(2), 218–232.

Cooper, H., Lindsay, J. J., Nye, B., & Greathouse, S. (1998). Relationships among attitudes about homework, amount of homework assigned and completed, and student achievement. *Journal of Educational Psychology, 90*(1), 70–83.

Cooper, H., Valentine, J. C., Nye, B., & Lindsay, J. J. (1999). Relationships between five after-school activities and academic achievement. *Journal of Educational Psychology, 91*(2), 369–378.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist, 30*(1), 1–14.

Cureton, L. W. (1971). The history of grading practices. *Measurement in Education, 2*(4), 1–8.

Davis, J. A. (1965). *Faculty perceptions of students: V. A. second-order structure for faculty characterizations* (College Entrance Examination Board RDR-64-5, No. 14; ETS RB-65-12). Princeton, NJ: Educational Testing Service.

Dewey, J. (1900). *School and society.* Chicago: University of Chicago Press.

Eccles (Parsons), J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). San Francisco: W. H. Freeman Company.

Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In N. Eisenberg (Ed.), *Social, emotional, and personality development* (Vol. 3, pp. 1017–1095). New York: John Wiley & Sons.

Ekstrom, R., Goertz, M., & Rock, D. (1988). *Education & American youth.* London: Falmer Press.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25*(4), 333–347.

Finn, J. D. (1993). *School engagement and students at risk.* Washington, DC: National Center for Education Statistics.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice, 12*(3), 23–30.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*(3), 193–202.

Frederiksen, N., Glaser, R., Lesgold, A. & Shafto, M. G. (1990). *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gamache, L. M., & Novick, M. R. (1985). Choice of variables and gender differentiated prediction within selected academic programs. *Journal of Educational Measurement, 22*(1), 53–70.

Gifford, B. R., & O'Connor, M. C. (Eds.). (1992). *Changing assessments: Alternate views of aptitude, achievement, and instruction.* Boston: Kluwer Academic Publishers.

Goldman, R. D., & Hewitt, B. N. (1975). Adaptation-level as an explanation for differential standards in college grading. *Journal of Educational Measurement, 12*(3), 149–161.

Goldman, R. D., & Slaughter, R. E. (1976). Why college grade point average is difficult to predict. *Journal of Educational Psychology, 68*(1), 9–14.

Goldman, R. D., & Widawski, M. H. (1976). A within-subjects technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement, 36*(2), 381–390.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley & Sons.

Hanks, M. P., & Eckland, B. K. (1976). Athletics and social participation in the educational attainment process. *Sociology of Education, 49*(4), 271–294.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research, 52*(1), 123–142.

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review, 102*(3), 458–489.

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

Humphreys, L. G. (1968). The fleeting nature of academic prediction. *Journal of Educational Psychology, 59*(5), 375–380.

Hunter, J. E., & Schmidt, F. L. (1982). *Meta analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage Publications.

Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H., & Frankel, M. R. (1994). *Second follow-up: Student component data file user's manual* (NCES 94–374). Washington, DC: National Center for Education Statistics, U. S. Department of Education.

Ingels, S. J., Dowd, K. L., Taylor, J. R., Bartot, V. H., Frankel, M. R., & Pulliam, P. A. (1995). *Second follow-up: Transcript component data file user's manual* (NCES 95–377). Washington, DC: National Center for Education Statistics, U. S. Department of Education.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books, Inc.

Keeton, M., & Associates. (1976). *Experiential learning: Rationale, characteristics, and assessment.* San Francisco: Jossey-Bass.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program. *Educational Measurement: Issues and Practices, 13*(3), 5–16.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II. Affective domain.* New York: David McKay Company, Inc.

Lamborn, S. D., Mounts, N. S., Steinberg, L., & Dornbusch, S. M. (1991). Patterns of competence and adjustment among adolescents from authoritative, authoritarian, indulgent, and neglectful families. *Child Development, 62*(5), 1049–1065.

Lemann, N. (1999). *The big test: The secret history of American meritocracy.* New York: Farrar, Straus, & Giroux.

Lewis, C., & Willingham, W. W. (1995). *The effects of sample restriction on gender differences* (ETS RR-95-13). Princeton, NJ: Educational Testing Service.

Lindquist, E. F. (1963). An evaluation of a technique for scaling high school grades to improve prediction of college success. *Educational and Psychological Measurement, 23*(4), 623–646.

Lindsay, P. (1982). The effect of high school size on student participation, satisfaction, and attendance. *Educational Evaluation and Policy Analysis, 4*(1), 57–65.

Linn, R. L. (1966). Grade adjustment for prediction of academic performance: A review. *Journal of Educational Measurement, 3*(4), 313–329.

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In Wigdor, A. K., & Garner, W. R. (Eds.), *Ability testing: Uses, controversies, and consequences* (Vol. 2, pp. 335–388). Washington, DC: National Academy Press.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*(1), 1–4.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295.

Marsh, H. W., (1991). Employment during high school: Character building or a subversion of academic goals? *Sociology of Education, 64*(3), 172–189.

Marsh, H. W. (1992a). Content specificity of relations between academic achievement and academic self-concept. *Journal of Educational Psychology, 84*(1), 35–42.

Marsh, H. W. (1992b). Extracurricular activities: Beneficial extension of the traditional curriculum or subversion of academic goals? *Journal of Educational Psychology, 84*(4), 553–562.

Marsh, H. W., & Yeung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89*(1), 41–54.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive.* New York: Appleton-Century-Crofts.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15*, 335–374.

Monk, D. H., & Haller, E. J. (1993). Predictors of high school academic course offerings: The role of school size. *American Educational Research Journal, 30*(1), 3–21.

Office of Civil Rights. (1999). *Nondiscrimination in high-stakes testing: A resource guide.* Washington, DC: U. S. Department of Education.

Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications.* Englewood Clifts, NJ: Prentice Hall.

Public Agenda. (2000, February 16). Reality check 2000 [Special Report]. *Education Week,* pp. S1–S8.

Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist, *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.

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

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.). *Changing assessments: Alternate views of aptitude, achievement, and instruction* (pp. 37–76). Boston: Kluwer Academic Publishers.

Richards, J. M., Jr., Holland, J. L. & Lutz, S. W. (1967). Prediction of student accomplishment in college. *Journal of Educational Psychology, 58*(6), 343–355.

Robinson, G. E., & Craver, J. M. (1989). *Assessing and grading student achievement.* Arlington, VA: Educational Research Service.

Rock, D. R., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for the NELS:88 base year through second follow-up.* (NCES 95-382). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Saslow, L. (1989, May 7). Schools say inflated grades cut grants. *New York Times,* p. 1.

Sewell, W. H., & Shah, V. P. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology, 73*(5), 559–572.

Shepard, L. A. (1992a). Commentary: What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternate views of aptitude, achievement, and instruction* (pp. 301–328). Boston: Kluwer Publishers.

Shepard, L. A. (1992b). Uses and abuses of testing. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (pp. 1477–1485). New York: Macmillan Publishing Company.

Shepard, L. A. (2000). *The role of classroom assessment in teaching and learning* (CSE Technical Rep. 517). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18*(9), pp. 8–14.

Starch, D., & Elliott, E. C. (1913). Reliability of grading work in mathematics. *School Review, 21*(5), 254–256.

Steinberg, J. (2000, December 22). Student failures causes states to retool testing programs. *New York Times,* p. 1.

Strenta, A. C., & Elliott, R. (1987). Differential grading standards revisited. *Journal of Educational Measurement, 24*(4), 281–291.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1991). *Sex differences in SAT predictions of college grades* (College Board Rep. No. 91-2, ETS RR-91-38). New York: College Entrance Examination Board.

Taber, T. D., & Hackman, J. D. (1976). Dimensions of undergraduate college performance. *Journal of Applied Psychology, 61*(5), 546–558.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research.* Newbury Park, CA: Sage Publications.

Werts, C., Linn, R. L., & Joreskog, K. G. (1978). Reliability of college grades from longitudinal data. *Educational and Psychological Measurement, 38*(1), 89–95.

Werts, C. E. (1967). The many faces of intelligence. *Journal of Educational Psychology, 58*(4), 198–204.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703–713.

Wilgoren, J. (2000, February 25). Cheating of statewide tests is reported in Massachusetts. *New York Times.*

Willingham, W. W. (1963). Adjusting college predictions of the basis of academic origins. In M. Katz (Ed.), *The twentieth yearbook of the National Council on Measurement in Education* (pp. 1–6). East Lansing, MI: National Council on Measurement in Education.

Willingham, W. W. (1965). The application blank as a predictive instrument. *College and University, Spring*, 271–281.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability.* New York: College Entrance Examination Board.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W. W., & Lewis, C. (1990). Institutional differences in prediction trends. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist, *Predicting college grades: An analysis of institutional trends over two decades* (pp. 141–160). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences* (ETS RR-00-15). Princeton, NJ: Educational Testing Service.

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement, 27*(2), 175–186.

Authors

WARREN W. WILLINGHAM has retired as Distinguished Research Scientist at Educational Testing Service in Princeton, NJ; warren131@aol.com. His research interests include test validity and fairness, policy issues in assessment, and admission to higher education.

JUDITH M. POLLACK is a director of research at ETS in Princeton, NJ. Her research interests include survey instrument and test design for large-scale assessments, longitudinol studies, and design and analysis of complex data sets.

CHARLES LEWIS is a professor of psychology at Fordham University and a Distinguished Presidential Appointee at ETS, Princeton, NJ. His research interests include test validity and fairness, test theory (including IRT and CBT), Bayesian methods, and hierarchical and generalized linear models.