Validity Decay in STEM and Non-STEM Fields of Study

Paul A. Westrick

WP-2014-05
July, 2014

ACT Working Paper Series

ACT®

Validity Decay in STEM and Non-STEM Fields of Study

Abstract: The purpose of this study was to determine if validity coefficients for ACT scores and high school grade point average (HSGPA) decayed or held stable over eight semesters of undergraduate study in science, technology, engineering, and mathematics (STEM) fields at civilian four-year institutions, and whether the decay patterns differed from those found in non-STEM fields at the same institutions.

Data from 62,122 students at 26 four-year institutions were analyzed in a hierarchical meta-analysis in which student major category (SMC), gender, and admission selectivity levels were considered potential moderators. Analyses for twelve subgroups were run. The results indicated that ACT score validity coefficients for STEM-Quantitative and STEM-Biological majors decayed less over eight semesters than the validity coefficients for non-STEM majors did. This was true for the uncorrected and corrected validity coefficients. For the HSGPA validity coefficients, this was true for the corrected validity coefficients.

Keywords: Validity decay, dynamic criterion, meta-analysis, STEM, ACT

Introduction

Past research has suggested that the validities of admission tests and high school class rank decay over eight semesters (Humphreys, 1968; Wilson, 1983). Early research suggested that the decay may be due to the students changing over time (Humphreys, 1968), though later research suggested that validities decay because the criterion changes over time (Humphreys and Taber, 1973; Powers, 1982). Butler and McCauley (1987), however, found validity stability instead of validity decay when using data from the United States Military Academy (USMA) and the United States Air Force Academy (USAFA). Among the reasons given for this stability were that, 1) at least half of the four-year curriculum at the military academies consisted of mathematics and natural science courses, 2) instructors teaching the same courses had to use common syllabi, and 3) the instructors at the academies tested the cadets more frequently than instructors tested students at civilian institutions. The first point is probably the most important one

because students taking courses in a program with a highly structured curriculum do not have the freedom to avoid difficult courses and select courses known for easy grading standards.

The Butler & McCauley (1987) results were unique in the validity decay literature, but the military academies are unique academic environments. Most students in the United States attend civilian institutions, but within civilian institutions the academic requirements in science, technology, engineering, and mathematics (STEM) fields are similar to those found in the military academies in that more than half the required courses are in mathematics and the natural sciences. Furthermore, STEM courses typically must be taken in sequential order, ensuring a high level of uniformity in the content of the prerequisite STEM courses that must be completed before attempting the advanced STEM courses. Lastly, and as discussed in greater depth below, the grading practices in STEM fields are more stringent than the grading practices in non-STEM fields (e.g., Goldman & Widawski, 1976; Strenta & Elliot, 1987). Although one cannot completely replicate the Butler & McCauley (1987) study while using data from civilian institutions, there would be important similarities with such a study and the results could be generalized to the much larger civilian college population.

The main purpose of this meta-analysis is to determine if validity coefficients for ACT scores, both composite scores and subject area test scores, and high school grade point average (HSGPA) decay over eight semesters of undergraduate study in STEM fields at civilian four-year institutions, and whether the decay pattern differs from that found in non-STEM fields at the same institutions. In this study, validity decay is defined as a general trend in which the correlations between a predictor variable and a criterion decrease each time the criterion is measured. With the passage of time, the validity coefficients trend downward, and the validity coefficient for the eighth semester is less than the validity coefficient for the first semester. Another question of interest is whether admission selectivity and gender further moderate the relationships between precollege predictors (ACT scores and HSGPA) and academic performance in the STEM fields.

Whereas most studies that focused on validity decay have used data from single institutions, this meta-analysis uses the data from 26 four-year institutions. This study draws upon and integrates research literature on the validity of traditional precollege academic predictors (admission test scores and HSGPA); validity decay; differential grading practices across fields of study; and characteristics of STEM fields and the students who enter these fields. The integration of this literature provided the rationale for a hierarchical moderator analysis in which the validity decay analyses were ultimately disaggregated by three student major categories (SMCs), two levels of institutional admission selectivity, and gender for a total of twelve subgroups.

Criterion Model of Validity

The word "validity" has meant different things to different people at different times. Brennan (2006) has traced the meaning of the word across the various editions of *Educational Measurement* and in a variety of contexts (e.g., K-12, licensure exams, admissions to higher education), and Kane (2006) has described the evolution of how the measurement community has conceptualized validity, from the criterion model, to the content model, to the construct model, and finally to his argument-based approach to validity. However, Kane noted that "for admissions, placement, and employment testing, the criterion model is still the preferred approach" (p. 17). With admissions, the focus is on predictive validity, and the outcome of the validity study, a correlation between the predictor variable and the criterion of interest, provides what may be seen as tangible evidence to support admission decisions. Institutions want to be able to predict future academic performance, defined as the grades students earn in the courses taught by the institutions' faculty members, and admission test scores and HSGPA typically provide the best information to make these predictions. For this reason, this study adheres to the criterion model with a focus on the predictive validity of ACT scores and HSGPA.

Validity studies generally report both observed and corrected correlations. It is well known that range restriction on the predictor variables and unreliability in the criterion reduce the size of observed

correlations (Cronbach, 1960; Gullicksen, 1987), and corrections for these artifacts can produce large differences between the observed correlations and corrected correlations. For example, Ramist et al. (1994, p. 13) reported uncorrected correlations between combined SAT scores (Verbal plus Math) and first-year GPA of .36. After correcting for range restriction, the correlation rose to .53, and then after correcting for unreliability in first-year GPA the fully corrected correlation rose to .57. Later studies (Bridgeman et al., 2008; Kobrin, Patterson, Shaw, Mattern, and Barbuti, 2008; Sackett, Kuncel, Arneson, Cooper, & Waters, 2009) have produced similar results after correcting for one or both of these artifacts. Corrected correlations between HSGPA and first-year GPA have ranged between .54 and .61, and corrected multiple correlations (test scores and HSGPA) have ranged between .62 and .68 (Bridgeman et al., 2008; Kobrin et al., 2008; Ramist et al., 1994).

Validity Decay

Wilson (1983) observed that most validity studies focused on the validity of test scores and HSGPA as predictors of first-year GPA and that research on their predictive validity beyond the first year was not as common. However, summarizing the research on this topic up through the early 1980s, he found that when predicting independently calculated GPA (e.g., first-semester GPA, second semester GPA, etc.), validity coefficients for precollege predictors such as test scores, high school rank (HSR), and HSGPA generally decayed over time. This trend was best described by Humphreys (1960) in his simplex model, in which the strongest relationships were between measurement variables that were chronologically closest to one another and weakened as the interval between measurements increased. For example, given GPAs for eight semesters, first semester GPA would have its highest correlation with second-semester GPA and its correlations with the other semester GPAs would monotonically decline to its lowest correlation, its correlation with eighth-semester GPA.

Of all the studies on validity decay, Humphreys' (1968) classic study stands out in the literature. Humphreys tracked the academic performance of students at the University of Illinois across eight

semesters and found that although ACT scores and HSR were valid predictors in every semester, they declined steadily over the four years. The correlations reported were actually averages of the correlations calculated by gender and college within the university, each weighted by *N*. This decline was true for the correlations calculated using all students enrolled in each semester (varying *N*s), when using just the data from only the students who had been continuously enrolled for eight semesters, and when using the data from the first analysis and then making corrections for range restriction with the continuously enrolled students. The explanation given for validity decay was that it was most likely that students were changing over time, hence the decline. Later, Humphreys and Taber (1973) used GRE scores to postdict undergraduate grades and found that the correlation coefficients were highest in the first semester and lowest in the eighth semester, a contradiction of the simplex model. One possible explanation they gave was that it the criterion, semester GPA, had changed between the freshman and senior years.

Wilson (1983), after reviewing the literature on the prediction of academic performance beyond the first year of college and his own research (1978, 1980, 1981), concluded that validity coefficients for test scores, HSGPA, and HSR declined over time and there appeared to be evidence that the decay was due to the criterion changing. Wilson's review of the literature was quite thorough, but other researchers continued to conduct research on validity decay. Powers (1982) found validity decay in the prediction of law school grades, but recognized that the first year of law school consists of common courses, the second year less so, and the third year is often filled with clinical courses, making comparisons more difficult. In their study on differential grading practices across fields of study, Elliott and Strenta (1988) found that adjusting GPAs for differential grading practices not only raised the validity coefficients but it decreased the amount of decay seen over time.

These two studies on course selection and differential grading practices helped explain why validity coefficients decayed over time, but they did not show or argue for validity stability. Rather, they provided different explanations or proposed methods to demonstrate that the decline in validity coefficients was not as steep as it initially appeared. A strong argument for validity stability came from

Butler and McCauley (1987) who used data from the USMA and at the USAFA. They found that the validity coefficients of SAT scores and high school class ranks did not decay over time at the USMA, and although they did decline at the USAFA, they did not decline nearly as much as Humphreys had found in his studies. As mentioned earlier, Butler and McCauley presented three possible factors for validity stability at the military academies, but probably the most important factor given was that the cadets at the military academies had little room to maneuver when it came to choosing courses. Civilian students experiencing academic difficulties are free to change to a major with less stringent grading practices and/or seek out instructors who are reputed to be easy graders (Prather Smith, & Kodras, 1979), and this mobility could explain why so many other studies reviewed by Wilson (1983) had found validity decay in academic settings. In a sense, by using only cadets at the military academies, Butler and McCauley largely controlled for differential grading practices across fields of study.

Though not dedicated to examining validity decay, a recent series of validity studies have looked at the validity of SAT scores and HSGPA as predictors of independently calculated annual GPA and cumulative GPA across four years (Kobrin et al., 2008; Mattern & Patterson, 2011a, 2011b, 2011c). Starting with data for 151,316 students at 110 institutions in the first year and finishing with 56,939 students at 55 institutions in the fourth year, within each institution they obtained the observed correlations and then made corrections for range restriction within each institution, using the 2006 cohort of SAT examinees (college bound seniors) as their reference population. They then obtained weighted average observed and corrected correlations for each predictor variable or combination of predictor variables. When looking at independently calculated annual GPAs, combining the results of the four studies showed a pattern of validity decay for both the observed and corrected validity coefficients for SAT scores and HSGPA. When looking at cumulative GPAs, the corrected and uncorrected validity coefficients for SAT scores and HSGPA held steady or increased slightly over four years.

In summation, the majority of research using independently calculated semester or annual GPA as the criteria has found that validity coefficients declined, or decayed, over consecutive terms or years, with

the Butler and McCauley's (1987) results standing out as an exception. The results of the key validity decay/stability studies are summarized in Table 1. In light of the Butler and McCauley study, the argument that validity coefficients do not necessarily decay over time when students are in a restricted curriculum that is heavy in mathematics and natural science courses must be taken as a distinct possibility.

STEM Fields

Although some consider the social sciences and psychology to be STEM fields (Green, 2007), a more restricted definition (Chen & Weko, 2009) limits inclusion to mathematics; natural sciences (including physical sciences and biological/agricultural sciences); engineering/engineering technologies; and computer/information sciences. These STEM fields have certain characteristics that distinguish them from non-STEM fields. One is that they require students to complete sequential courses in mathematics and the natural sciences (Kokkelenberg & Sinha, 2010; Oh, 1976; Ost, 2010; Prather & Smith, 1976), which was a key factor in the Butler and McCauley (1987) study. Though the expression "two cultures" is commonly associated with Snow's (1959) lecture on the differences between the humanities and the sciences, Elliott and Strenta (1988, p. 334) used it to describe the curricular differences between the areas of study, with mathematics and science curriculums being "hierarchically organized and unforgiving of any lack in basic knowledge or skill." Students cannot opt out of these sequential courses if they want to remain in a STEM field and subsequently earn a degree in the field. Being able to complete the courses at the end of the program requires students to use the knowledge that they had acquired in the preceding courses, which go back to the initial mathematics and natural science courses completed in the first year of study. Success in those initial courses requires sufficient academic preparation before college, so it is imperative that aspiring STEM students enter college prepared for the academic work.

Another defining characteristic of STEM fields is that the grading standards appear to be more stringent (Elliott & Strenta, 1988; Goldman et al., 1974; Goldman & Hewitt, 1975, 1976; Goldman &

Widawski, 1976; Hewitt & Jacobs, 1978; Oh, 1976; Prather & Smith, 1976; Prather et al., 1979; Strenta & Elliott, 1987; Strenta, Elliott, Adair, Matier, & Scott, 1994). Some researchers have developed grade adjustment methods at the individual course level (Berry & Sackett, 2009; Noble & Sawyer, 1987; Stricker, Rock, & Burton, 1993; Young, 1990a, 1990b, 1993), while others have tried to develop grade adjustment methods to make GPAs for students in different majors more comparable (e.g., Goldman & Widawski, 1976; Strenta & Elliot, 1987; Elliott & Strenta, 1988; Pennock-Roman, 1994). Distinctions between quantitative and non-quantitative fields have been made, though Pennock-Roman reported that the grading standards in biological fields did not fit into this dichotomy. Students also seem to be aware of these different grading standards (Goldman & Hewitt, 1975; Hewitt & Jacobs, 1978), which influences their selection of courses and majors. These differences in grading practices across fields may help explain who enters and succeeds in STEM fields.

Ability Level of STEM Majors

Four decades ago, Burnham and Hewitt (1972) concluded that, among the students at Yale, those who had high verbal aptitude test scores and high mathematics achievement test scores were the only students who really had a choice between fields of study. More specifically, they suggested that mathematics was what separated students. They argued that schools ought to require students to take College Entrance Examination Board achievement tests in mathematics and the natural sciences and select students who scored high on these tests as well as on the verbal tests. Then the schools would have incoming students "who are in fact free to choose their prospective major on the basis of positive interest (p. 410)." What they avoided saying directly was that those who lacked mathematical and science ability had no choice over whether to be a STEM major or a non-STEM major. They had to be non-STEM majors.

It was a somber message, over the following decades other researchers have made similar arguments. The general message has been that as the STEM fields require sequential coursework that

cannot be avoided, and as the grading practices are stricter in STEM fields, students seem to screen themselves into or out of STEM fields on their own. Students enrolling in STEM programs generally have higher levels of precollege academic preparation as measured by HSGPA and admission test scores than do students who enrolled in non-STEM programs (Elliott & Strenta, 1988; Green, 1989; Nicholls, Wolfe, Besterfield-Sacre, Shuman, & Larpkiattaworn, 2007; Ost, 2010; Pennock-Roman, 1994; Price, 2010; Strenta & Elliot, 1987; Strenta et al., 1994; White, 1992). This was an important factor in the series of studies on differential grading practices at universities in California (Goldman et al., 1974; Goldman & Hewitt, 1975, 1976; Goldman & Widawski, 1976) where students in the sciences, mathematics, and engineering consistently had HSGPAs and SAT scores, especially SAT-Mathematics scores, that were higher than those associated with students in other majors. Strenta and Elliott (1987; Elliott & Strenta, 1988; Strenta et al., 1994) also found that students in the sciences, mathematics, and engineering had higher HSGPAs and SAT scores than their non-STEM peers, and much as Goldman and Hewitt (1976) had observed, it was mathematics ability that set the science, mathematics, and engineering students apart from students in other majors. More recent research found that despite making up only 22.8% of the first-year students, STEM majors were 31.1% of the entering students who had earned at least a B average for their HSGPA, and they were 51.1% of the entering students who had scored in the top quartile on their admission tests (Chen & Weko, 2009). STEM majors were also more likely to enroll in highly selective institutions (32.6% versus 21.1%) than their non-STEM counterparts.

Gender Differences in STEM Fields

Female representation in the STEM fields has been an issue of interest for decades (Wai, Cacchio, Putallaz, & Makel, 2010), with researchers providing multiple explanations for why males are more prevalent than females are in certain STEM fields (e.g., Ceci & Williams, 2007), especially the more quantitative fields. As Kimura (2007) pointed out, even though work in the biological sciences – a field where females are well-represented – require mathematics, the level of mathematics required is not as high as that required for physics, where females are less represented. However, even among youth

identified as being mathematically precocious (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000), in a twenty-year follow-up study, males were nearly twice as likely as females to have earned a bachelor's degree in mathematics or the inorganic sciences, and females were nearly twice as likely as males to have earned a bachelor's degree in the life sciences or humanities. Looking at national data for all STEM majors, in 2007 males earned roughly 80% of all the bachelor's degrees awarded in engineering, computer sciences, and physics, and females earned 60% of all bachelor's degrees in biological sciences and 50% of all bachelor's degrees in chemistry (National Science Board, 2010). Another possible explanation for these gaps is that among females with high mathematical abilities, many are even stronger in other areas, and students tend to follow academic and career paths that match their strengths (Lubinski & Benbow, 2007). Attempting to explain why males and females enter different STEM fields at different rates is beyond the scope of this paper, but the differences in male and female participation rates across STEM fields provides further support for distinguishing between STEM-Quantitative and STEM-Biological majors. Furthermore, differential validity research has consistently found that females earn higher grades and that the validity coefficients for females are slightly higher than those for males (Young & Kobrin, 2001; Zwick, 2006). Considering these gender differences in participation rates, grades earned, and validity coefficients, gender should be considered as a potential moderator in validity decay analyses.

Institutional Admission Selectivity

Past research suggests that institutional admission selectivity should be considered a potential moderator of validity coefficients. Regarding admission selectivity, Kobrin et al. (2008) found that validity coefficients for SAT scores were higher at more selective institutions than they were at less-selective institutions, but validity coefficients for HSGPA were highest at the less-selective institutions. This trend continued in subsequent research over the next three academic years (Mattern & Patterson, 2011a; 2011b; 2011c). Allen and Robbins (2010) found that validity coefficients for ACT Composite scores and HSGPA were higher at four-year institutions, most having admission policies in which most students came from the upper half of their high school classes, than at two-year institutions, most of

which had open admissions. Given these findings, admission selectivity should also be considered a potential moderator of validity coefficients over time.

Hypothesis

A fundamental concept to remember in correlational studies is that analyses conducted on a group that contains subgroups with different means and standard deviations on any of the variables may lead to spurious correlations (Kirk, 1999). STEM majors have higher mean admission tests scores and HSGPAs (Chen & Weko, 2009), hence they should not be analyzed together with non-STEM majors. Similarly, students at more-selective institutions have higher mean admission tests scores and HSGPAs than students at less-selective institutions. Finally, males tend to have slightly higher mean admission test scores and more variance in their scores, whereas females tend to have higher mean HSGPAs (Kobrin & Patterson, 2011a; 2011b; 2011c). Given these differences, the validity coefficients from the overall results may be misleading. However, as the results presented by Humphreys (1968) and Butler and McCauley (1987) did not make distinctions between academic majors, it was decided to first meta-analyze the data without considering potential moderators. Subsequent analyses included SMC, gender, and institutional admission selectivity levels as potential moderators.

Hypothesis: Regardless of gender and institutional admission selectivity level, student major category will moderate the relationships between precollege academic predictors – ACT scores and HSGPA – and independently calculated semester GPA across eight consecutive semesters. Within each gender by admission selectivity grouping, students in the STEM-Quantitative and STEM-Biological categories will have less validity decay than the students in the non-STEM categories.

Methodological Issues

Predictions of future behavior will always be less than perfect (Thorndike, 1963), and a number of factors cause this imperfection, including measurement error in the predictors and the criterion; heterogeneity in the criterion; the limited scope of single predictors; group differences; and the impact of

individual experiences between the time the predictor measurement was taken and the time the criterion measurement was taken. Another artifact to add to the list is range restriction, which is widely known to reduce the magnitude of correlations.

Many of these issues were directly addressed in this meta-analytic study. Validity studies from 26 institutions were combined using the Hunter and Schmidt (2004) meta-analytic techniques, which use a random effects model that allows validity coefficients to vary across institutions while also determining estimated mean correlations between the predictors and the criterion. This methodology makes corrections for measurement error and range restriction at each institution before pooling data across institutions. This pooling also diminishes the problem of sampling error, which is common in individual, single institution studies with small sample sizes.

In addition to correcting for measurement error and range restriction, a hierarchical moderator analysis addresses the problems arising from heterogeneity and group differences. Heterogeneity in the criterion, which concerns differences in the "common" criterion across institutions, fields of study, and teachers, was addressed by grouping students according to fields of study and by admission selectivity. The only group difference addressed in this study was gender. Though very much a grouping of interest, race/ethnicity was not analyzed as a moderator because of insufficient numbers for some of the groups across institutions and fields of study.

Finally, this validity decay study addressed the issue of the time between when the predictor measurement was taken and when the criterion measurement was taken. Finding validity stability, as Butler and McCauley (1987) found, would suggest that individual experiences between the time that the predictor measurement was taken and the time that the criterion measurement was taken have little if any impact on the relationship between the predictor and the criterion.

Methods

Sample

Data for the current study came from 26 four-year institutions that had participated in various ACT research services or partnerships. The institutions were located in 13 states, mostly in the Midwest and South, and of the 26 institutions, 23 were public and three were private. Admission selectivity was defined in accordance with the classification system utilized by ACT (2010) and summarized in Table 2. In this data set, one institution was classified as highly selective; nine were classified as selective; 15 were classified as traditional; and one was classified as open. No institutions in the data set fell into the liberal classification level. As there were not enough institutions at each level, the highly selective and selective institutions were grouped as "more selective" institutions, and the traditional, liberal and open admission institutions were grouped as "less selective" institutions.

Inclusion in the study depended on meeting criteria at the institution and individual levels. To be consistent with the Humphreys (1968) and Butler and McCauley (1987) studies, students had to be continuously enrolled in the same four-year institution. Therefore, at the institution level, schools had to have at least four years of follow-up data on their students, to include semester GPA for eight consecutive semesters, from the first semester of the first year through the second semester of the fourth year. Cohort years between 2000 and 2005 were included. To ensure that students were continuously enrolled at the same institution, all dropouts and transfer students were excluded, as were students who dropped out and later returned to the same institution.

A fundamental goal of the study is to compare three student major categories (SMCs; STEM-Quantitative, STEM-Biological, and non-STEM) within institutions because institutions that offer all three options give students a choice of entering one of the STEM areas of interest as well as any of the numerous non-STEM fields. Therefore, institutions that did not have at least three observations in each of the six gender by SMC subgroups were excluded. This liberal minimum requirement of having at least three observations in each subgroup made it possible to include most of the less-selective institutions that had few STEM majors but many more non-STEM majors. At the individual student level, students had to

have valid ACT scores, HSGPA, a semester GPA for each semester, and a cumulative GPA for each semester.

An additional requirement was that students had to have a Classification of Instructional Programs (CIP) code (National Center for Educational Statistics, NCES, 2002) associated with their records. The CIP code is a six-digit number that identifies a student's declared major. The first two digits are the most general categorization, the first four digits provide an intermediate categorization, and the full six digits provide the most specific categorization. In light of the findings of Pennock-Roman (1994), in this study students with a CIP code of 11 (Computer Sciences), 14 (Engineering), 27 (Mathematics and Statistics), and 40 (Physical Sciences, primarily physics and chemistry) were pooled to create the STEM-Quantitative category, and students with a CIP code of 26 (Biological and Biomedical Sciences) were used to create the STEM-Biological category. All other students with a declared major were classified as non-STEM.

As grading practices differ across majors, and these differences potentially moderate the relationship between precollege predictors (ACT scores and HSGPA) and undergraduate GPA, students who changed SMCs while at the same institution were excluded from the analyses. However, students were allowed to change majors within their SMC. For example, a student who initially majored in chemistry and later changed to engineering would have been included in the study because both majors (CIP codes 40 and 14) would fall into the STEM-Quantitative category. If the student had changed his/her major from chemistry to communications, the student would have changed categories (STEM-Quantitative to non-STEM) and would have been excluded.

The data set consisted of ACT-tested students who enrolled as first-time students entering in the fall term from 2000 to 2005, a total of up to six cohorts per institution. Few institutions had enough students to conduct moderator analyses based on racial/ethnic categories, but the overall racial/ethnic breakdown was as follows: 48,949 Caucasians, 3,868 African-Americans, 2,638 Hispanics/Latinos, 1,695 Native

Americans/Native Alaskans, and 1,857 Asian-Americans. A total of 611 students identified themselves as Multiracial, and 513 identified themselves as "Other". A total of 1,991 students either selected "Prefer not to respond" or simply did not answer the item.

Validity generalization studies seek not to simply describe how the results apply to the study participants but to the overall population of interest from which the sample was taken. Data from this population, or referent group, are needed to make corrections for range restriction, which is discussed in detail later. This population of interest is the national ACT examinee population, to which the results of this study will be generalized. To match the cohort years described earlier, data from all 1999-2005 ACT examines were used as the referent populations[1]. For the analyses is this study, the two referent groups were the male and female ACT-tested populations. Table 3 contains descriptive statistics for these referent populations and the overall ACT-tested population. Descriptive statistics for the overall group and the subgroups for the moderator analyses in this study are presented in Table 4. Consistent with the literature, the STEM majors had higher mean ACT scores and HSGPAs than those of the non-STEM majors, both overall and within institutional admission selectivity levels.

Independent Variables: ACT Scores and HSGPA

The ACT® college readiness assessment includes four multiple-choice subject area tests – English, Reading, Mathematics, and Science – and an optional Writing Test (ACT, 2007). Raw scores for each subject area test are converted to scale scores that range from 1 to 36, and the Composite score also ranges from 1 to 36. The ACT tests were developed by content experts who consulted junior and senior high school state standards for the content areas (grades 7 through 12), textbooks on state-approved lists, high school teachers, and college and university faculty members to ensure a high degree of agreement with what the students are studying in junior and senior high school and what they need to know to succeed in entry-level courses in college. The reliability of the subject area test scores and the Composite

score reported in the ACT Technical Manual (ACT, 2007) are as follows: English, .91; Mathematics, .91; Reading, .85; Science, .80; and the Composite, .96.

The measure of HSGPA in this study was based on students' self-reported high schools grades in the four core subject areas: English, mathematics, social science, and natural science. Although students report grades on up to 30 high school courses, only grades earned for the first 23 core subject area courses are used in calculating HSGPA. ACT research found that the median correlation between self-reported high school grades and actual grades on transcripts was .79 (Schiel & Noble, 1991). This median correlation was used as the reliability estimate of HSGPA. Note that this estimate came from a restricted sample, and it was used as the restricted reliability estimate for each institution. Given the estimated reliability of HSGPA in the restricted population at each school (.79), individual corrections for range restriction at each institution in this study were made by applying equation 3.17c from Hunter and Schmidt (2004) so that each institution had an institution-specific estimate for the reliability of HSGPA in the unrestricted national population. This was the opposite of what was done for ACT scores, where the reliability estimate for the national, unrestricted population was known and the institution-specific estimate for the reliability of HSGPA in the restricted school population had to be estimated.

Dependent Variable: Semester GPA

The dependent variable in the validity analyses was independently calculated semester GPA as reported by the institutions. All institutions reported semester GPA on a four-point scale. The reliability of GPA for a full academic year is typically calculated by using the correlation between semester GPAs within an academic year and applying the Spearman-Brown Prophecy Formula. However, as the prediction of semester GPA is the objective of this study, the reliability of semester GPA was estimated by using the correlation between adjacent semesters within an academic year. An alternative approach would have been to use an average of the two correlations available between adjacent semesters, but this would have worked only for semesters two through six, as semesters one and eight have only one adjacent

semester. Furthermore, the correlations between adjacent semesters within an academic year in Humphreys' (1968) validity decay study were higher than the correlations between semesters separated by a summer break, resulting in an up-down pattern over eight semesters and suggesting that the correlations between semesters within an academic year were better representations of the relationship. For this reason, only the correlations between adjacent semesters within an academic year were used.

Students who graduated in less than eight semesters were included in the study (n=3,956). For students who graduated at the end of the fifth, sixth, or seventh semester, their cumulative GPAs were used as their semester GPAs over the following semester(s). These proxy semester GPAs were included in the estimation of the reliability of semester GPA.

Moderator Variables

This study has three hypothesized moderator variables: admission selectivity, gender, and SMC. The moderator of most interest in this study is SMC. Some majors are more popular than others, and not every major is offered at each institution, so making comparisons at the level of the four-digit or six-digit CIP code was impractical. As described earlier, three SMCs were created: STEM-Quantitative, STEM-Biological, and non-STEM. Students who did not have a valid CIP code for each semester were excluded because they could have been in any of the three categories. Students were further subdivided by gender and institution admission selectivity, described earlier.

Meta-Analytic Techniques

The Hunter and Schmidt (2004) meta-analytic techniques were used to analyze the data. This methodology permits corrections for sampling error, measurement error, and range restriction. As institutions generally do not make admission decisions strictly using a top-down approach based upon test scores and HSGPA, corrections were made for indirect range restriction. Note that the methodology

makes corrections to estimate the correlation between the true scores on the predictor (T) and the true scores on the criterion (performance, P), ρ_{TP}. When the objective is the estimation of the mean operational validities of the predictor measures, as in this study, measurement error is reintroduced in T at the institutional study level before meta-analyzing the results (Hunter, Schmidt, & Le, 2006; Le & Schmidt, 2006).

As noted earlier, most validity decay/stability studies were conducted at single institutions (e.g., Butler & McCauley, 1987; Elliot & Strenta, 1988; Humphreys, 1968; Wilson, 1978, 1980; Wilson, 1978, 1981). The pooling of research results from individual institutions into a single meta-analysis allows researchers to generalize the results to institutions not included in the meta-analysis. Though the focus of this study was on validity decay seen in the estimated mean correlations that were calculated for each relationship, a 90% credibility interval was calculated for each corrected validity coefficient, each providing a range where we would expect to see the parameter values for 90% of all institutions to fall. The 90% credibility intervals (and 95% confidence intervals) that were calculated for each of the six predictor-criterion relationships across eight semesters for each of the subgroups are not reported here but are available upon request.

Hierarchical Moderator Analysis

Conducting multiple moderator analyses on moderators one by one can be misleading because the moderators may be correlated (Hunter & Schmidt, 2004). Therefore, a hierarchical meta-analysis was conducted with observations disaggregated into twelve subgroups by SMC, gender, and institutional admission selectivity.

Results

The results for the validity meta-analyses for ACT Composite (ACTC) scores, ACT English (ACTE) scores, ACT Mathematics (ACTM) scores, ACT Reading (ACTR) scores, ACT Science (ACTS)

scores, and HSGPA are presented in Tables 5 through 10. Corrected and uncorrected validity coefficients rounded to the second decimal place are reported across eight semesters. On the right side of each table are two columns, one showing the amount of change in the validity coefficients between the first and eighth semesters and the other showing the percent changed. The percent changed is the amount of change before rounding divided by the unrounded validity coefficient for the first semester.

Hierarchical Moderator Analysis Results

The hypothesis stated that student major category, admission selectivity, and gender would jointly moderate the relationships between precollege academic predictors – ACT scores and HSGPA – and independently calculated semester GPA across eight consecutive semesters. Students in the STEM-Quantitative and STEM-Biological categories would have less validity decay than the students in the non-STEM categories within their respective gender by admissions selectivity categories. The upper halves of the tables contain the observed correlations, and the bottom halves contain the corrected correlations. As range restriction and measurement error distort the relationship between the predictors and the criteria, the results for the corrected correlations are emphasized. Note that the corrections for range restriction and measurement error increased the validity coefficients for ACT scores and HSGPA, though more so for HSGPA due to greater amounts of range restriction and lower reliability estimates for HSGPA. Further note that with the increases in the corrected validity coefficients, the amount of change between the first and eighth semesters also increased. However, when looking at the percentage of change in the final column, the results for the corrected validity coefficients are quite similar to those for the uncorrected validity coefficients.

Females at more-selective institutions.

Among the females at schools with more-selective admission standards, the STEM-Quantitative and STEM-Biological majors had less validity decay than the non-STEM majors. For the ACTC-GPA relationships (Table 5), the amount of decay in the corrected correlations was -.07 for the STEM-

Quantitative majors, -.15 for the STEM-Biological majors, and -.29 for the non-STEM majors. For the ACT subject-area tests (Table 6, ACTE; Table 7, ACTM; Table 8, ACTR; Table 9, ACTS), the same patterns held. The STEM-Quantitative majors had the least amount of validity decay, followed by the STEM-Biological majors, and then followed by the non-STEM majors, who had the most validity decay. For the corrected HSGPA-GPA relationships, the amount of decay was only -.02 for the STEM-Quantitative majors. A larger amount of decay was associated with the validity coefficients for the STEM-Biological majors (-.19, -21.9%), but the non-STEM majors still had the largest amount of decay (-.22; -29.9%).

Females at less-selective institutions.

Among the females at schools with less-selective admission standards, the results for both the ACT-GPA relationships (Table 5 through Table 9) and the HSGPA-GPA relationships (Table 10) indicate that both the STEM-Quantitative and STEM-Biological majors had less validity decay over eight semesters than the non-STEM majors had, at least when considering the amount of change. In one instance the STEM-Quantitative majors had a larger percentage of change in the corrected correlations; in Table 9 (ACTS) the amount of change was -.19 for the STEM-Quantitative majors and -.22 for the non-STEM majors, but the percentage of change was -38.7% for the STEM-Quantitative majors and -37.8% for the non-STEM majors. Aside from this decline, the STEM-Quantitative majors actually had higher validity coefficients in the eighth semester than they had in the first semester for all the other ACT-GPA relationships, with increases ranging from .01 (ACTE) to .14 (ACTR). That is, they had validity growth rather than validity decay. Keep in mind that this group is the smallest of the twelve subgroups (n=174, k=16), and the results should be interpreted with caution.

Males at more-selective institutions.

Among the males at more-selective institutions, the STEM-Biological majors had the least amount of validity decay for the ACT-GPA relationships (Table 5 through Table 9). Following the

STEM-Biological majors were the STEM-Quantitative majors, and the non-STEM majors had the largest amounts of validity decay, though the amount of difference between the non-STEM and STEM-Quantitative subgroups were often small, ranging between .03 and .07 for the corrected correlations. For the HSGPA-GPA relationships (Table 10), the non-STEM majors again had the largest amount of validity decay over eight semesters (-.24, -34%).

Males at less-selective institutions.

Among the males at schools with less-selective admission standards, the results for both the ACT-GPA relationships (Table 5 through Table 9) and the HSGPA-GPA relationships (Table 10) indicate that both the STEM-Quantitative and STEM-Biological majors had less validity decay over eight semesters than the non-STEM majors had. The STEM-Biological majors actually had validity growth for three of the five ACT-GPA relationships, ranging from .01 (ACTC) to .05 (ACTE), and validity growth for the HSGPA-GPA relationship (.03) in Table 10. As with results for the female, STEM-Quantitative majors at less-selective institutions, this was a small subgroup (n=194, k=16), and the results should be interpreted with caution.

Validity generalization output

Though not reported here due to space limitations, 90% credibility intervals were calculated for each of the 624 corrected validity coefficients. For the overall analyses (Table5), none of the credibility intervals contained zero, indicating that there were purely positive relationships between precollege academic achievement measures (ACT scores and HSGPA) and undergraduate GPA across four years. Ten of the twelve subgroups in the final hierarchical moderator analysis also had 90% credibility intervals that did not contain zero. However, for the two smallest subgroups, female, STEM-Quantitative majors at less-selective institutions and male, STEM-Biological students at less-selective institutions, some of the credibility intervals did contain zero. For the female, STEM-Quantitative majors at less-selective institutions, six of the 48 credibility intervals contained zero: two for ACTR scores, two for ACTS scores,

and two for HSGPA. For the male, STEM-Biological majors at less-selective institutions, 14 of the 48 credibility intervals contained zero: one for ACTC scores, one for ACTM scores, seven for ACTR scores, and five for ACTS scores.

Discussion

Validity decay/stability has been a somewhat neglected area of research for the better part of two decades. The Humphreys (1968) study and the Butler and McCauley (1987) study were highlighted because the Humphreys study has been the prime example of validity decay and the Butler and McCauley study has been the prime example of validity stability. A third study of interest, conducted by Pennock-Roman (1994), did not pertain to the validity decay/stability debate, but it served as a catalyst for splitting the STEM majors into the STEM-Quantitative and STEM-Biological categories.

The main objective of this study was to determine if validity decay could be minimized by separating STEM majors from non-STEM majors. The inspiration for this came from Butler and McCauley's (1987) study, in which the researchers found validity stability using data from the USMA, which had highly structured curriculums that required cadets had to take more than half of their courses in mathematics and the sciences. As the majority of courses taken by STEM majors are also in mathematics and science, and these courses are typically completed in a structured, sequential order, it was anticipated that validity stability could be found or that at least the amount of validity decay would be less than that found in previous studies (e.g., Humphreys, 1968). Pennock-Roman's (1994) insights on differential grading and her observation that the grading profiles for the biological sciences did not fit with either the quantitative or non-quantitative fields led to the decision to split STEM majors into two categories, STEM-Quantitative and STEM-Biological, with all other majors classified as non-STEM majors. Based on the literature on differential validity, these three subgroups were further subdivided by gender and two levels of institutional admission selectivity for a total of twelve subgroups. The validity coefficients for

ACT scores and HSGPA across eight semesters from 26 four-year institutions were then meta-analyzed in three analyses.

The results supported the hypothesis. Running separate analyses for the STEM and non-STEM majors (Tables 5-10) helped reduce the amount of validity decay seen over eight semesters for the STEM majors. For ACT scores, when considering the amount of change from the first semester to the eighth semester, the corrected validity coefficients for the STEM-Quantitative and STEM-Biological majors always declined less than the corrected validity coefficients for the non-STEM majors did. When considering the percentage of change, it was true all but once. For the HSGPA-GPA relationships, the amount and percentage of validity decay in the corrected validity coefficients were always less for the STEM majors than they were for the non-STEM majors. In general, there was less validity decay associated with females STEM-Quantitative majors than there was with male STEM-Quantitative majors, though closer analysis found that the credibility intervals for both groups overlapped considerably.

As noted above, the results for the two smallest groupings – female, STEM-Quantitative majors at less-selective institutions (n=174), and male, STEM-Biological majors at less-selective institutions (n=194) – were the only two groups to show validity growth over eight semesters, at least for ACT scores. However, it is also worth noting that between the first and eighth semesters are extreme peaks and valleys. For example, for the ACTC-GPA relationships for female, STEM-Quantitative majors at less-selective institutions (Table 5), the corrected correlations ranged from .58 to .62 for six of the semesters, but rose to .65 and .71 in the third and fifth semesters, respectively. For the ACTC-GPA relationships for male, STEM-Biological majors at less-selective institutions (Table 5), although the corrected correlations for six of the semesters ranged between .43 and .46, in the fourth and sixth semesters they dropped to .26 and .21, respectively. Similar patterns can be seen for these two STEM groups for the ACT subject area tests and HSGPA. The presence of these peaks and valleys for these two groups are similar to those seen at the USAFA in the Butler and McCauley (1987; see Table 1) for the SAT-Verbal and high school rank.

For the remaining ten groups in the final analyses in this study, the pattern was validity decay across eight semesters for ACT scores and HSGPA.

The results suggest that validity decay is related to how observations are grouped. As discussed earlier, different fields of study appear to have different grading standards (e.g., Elliott & Strenta, 1988), but another factor to consider is that pooling different subgroups together may lead to spurious effects when the subgroups having different means or standard deviations on the predictors and criteria (Kirk, 1999). In this study the STEM majors were separated from the non-STEM majors, and in Table 4 it can be seen that within the admission selectivity by gender breakouts, the three SMCs had different means and standard deviations. Although the STEM-Quantitative category included four two-digit CIP families, the non-STEM category included 39 CIP families, a much more diverse mixture of students. While the STEM fields provided a good starting point, future analyses should be conducted by splitting apart the non-STEM majors. Perhaps the amount of validity decay in the non-STEM groups can be reduced by creating smaller subgroups for students majoring in areas such as education, business, the social sciences, and the humanities.

One goal of this study was to conduct a hierarchical moderator analysis with observations broken out by student major categories, gender, and admission selectivity so that any possible interactions could be identified. However, breaking out the observations by three potential moderators resulted in two subgroups – female, STEM-Quantitative majors at less-selective schools, and male, STEM-Biological majors at less-selective schools – that had very few observations spread across sixteen institutions. By setting the minimum number of observations at three for each subgrouping, a number of negative correlations led to high levels of variance and the 90% credibility intervals contained zero. While the hierarchical moderator analysis may have been a step too far, especially for the STEM groups at less-selective institutions, it did provide useful information. As noted earlier, the results suggested that among the STEM-Quantitative majors there was generally less validity decay for the females than there was for the males.

Limitations and Future Research

Although this study has provided a thorough examination of validity decay and validity stability, the study also has a number of limitations. While demonstrating that validity decay can be reduced by stratifying on academic field of study and other relevant variables, such as gender and admission selectivity, the results did not match the validity stability results found by Butler and McCauley (1987). Only the two smallest STEM groups in this study show validity stability or growth, but they also were the only two groups to have 90% credibility intervals that contained zero. In future research, it would be wise to raise the minimum number of observations. Keep in mind that the institutions included in this study were a reflection of the institutions willing to collaborate with ACT for at least four years and provide all the needed variables in every semester. Having more institutions with larger sample sizes would have been advantageous.

A central finding of this study was that by isolating the STEM-Quantitative and STEM-Biological majors, the validity coefficients for these subgroups showed less decay over time and in some cases the validity coefficients actually increased. When subgroups with different means or standard deviations are combined together, correlations for this combined group may be misleading. This was the point of the hierarchical moderator analyses. However, this study has not demonstrated that validity decay is inevitable for non-STEM majors. Within the non-STEM category are multitudes of different fields of study, each of which may also show less validity decay and possibly validity stability or growth if separated from the overall group. Future research should examine the possibility of validity stability among disaggregated non-STEM majors.

Consistent with previous validity decay/stability studies (Butler & McCauley, 1987; Humphreys; 1968), this study included only students who were continuously enrolled over eight semesters. This study went a step further by ensuring that students had to be in the same SMC over the eight consecutive semesters. This made the analyses easier because the number of observations remained constant in each

semester. It also made the results easier to interpret, as there was never a question as to how long the students were in a given category, or how many times they switched groups, or whether they had stopped out of school and returned. The downside of this approach is that in reality students do switch majors, and they do drop out, and some stop out for a semester or more and then return to school. Additional research on academic retention and the migration of students into and out of the STEM fields is needed.

Finally, this study has laid a foundation for future research on the gender gap in the STEM-Quantitative fields. It is interesting to observe that within their admission selectivity levels, males and females in the same SMC were more alike than with same sex members of other SMCs (Table 3). For example, female STEM-Quantitative majors had ACT score and HSGPA profiles that were more similar to the male STEM-Quantitative majors than they were to their female STEM-Biological and female non-STEM majors. When looking at ACT Composite scores and HSGPA, the gender differences for both STEM-Quantitative majors and STEM-Biological majors were small. However, within both STEM groups at both levels of admission selectivity males tended to have higher mean ACTM and ACTS scores and females tended to have higher mean ACTE and ACTR scores.

Given the smaller differences in ACTC scores, it may be that the larger differences in the ACT subject area scores are due to males and females investing their cognitive resources more heavily in subjects that match their interests (Cattell, 1987). Research that integrates the students' responses to the ACT Interest Inventory (ACT, 2009) with their ACT scores and HSGPAs is currently underway. Future gender-focused STEM research should also try to integrate additional noncognitive measures, as previous research has found that female students score slightly higher on scales of noncognitive skills in the areas of academic discipline, commitment to college, and study skills (Allen et al., 2008; Le, Casillas, Robbins, & Langley, 2005), and gender differences in motivation are related to gender differences in timely degree completion (Allen & Robbins, 2010). While interests may be more important when examining gender differences in entering the STEM-Quantitative and STEM-Biological fields, these non-cognitive factors

may also be related to gender differences in retention and degree completion, another area in need of more research.

Conclusion

This study has made a number of significant contributions to the literature. First, this meta-analysis was one of the largest validity decay studies conducted to date, including more than 60,000 students from 26 four-year institutions. The size of the sample made it possible to conduct a hierarchical moderator analysis, and the results out to the eighth semester indicate that there are differences in validity coefficients for different subgroups, especially when looking across student major categories.

A second contribution is that the results support the theory of a changing or dynamic criterion. The idea that the criterion changes over time is not new, and a number of researchers have studied and debated the idea of the changing or dynamic criterion, especially in the employee selection literature (e.g., Austin, Humphreys, & Hulin, 1989; Barrett & Alexander, 1989; Barrett, Alexander, & Doverspike, 1992; Barrett, Caldwell, & Alexander, 1985; Humphreys, 1976; Mauger & Kolmodin, 1975; Steele-Johnson, Osburn, & Pieper, 2000). An attempt was made to replicate the Butler and McCauley (1987) study by separating STEM students from non-STEM students at civilian schools, which would create a subpopulation that was similar to the STEM students used in their study, cadets at the military academies, notably the USMA. One possible reason that the results of the STEM majors in this study differed from those found at the USMA (Butler & McCauley, 1987) is that the criterion changed more at the civilian schools than it did at the USMA. A defining characteristic of the USMA was the highly structured curriculum with a common core of courses, which meant that they probably had fewer electives to choose from over four years than their counterparts at civilian institutions had. In this study, it was hoped that separating the STEM majors from the non-STEM majors would create subgroupings that faced a highly structured curriculum similar to that found at the USMA. However, the civilian STEM majors in this study probably had much more latitude in selecting electives both within and outside of their academic

fields than the cadets at the USMA had, especially in the last two years of their academic programs. Consequently, the criterion may have changed more for the STEM majors in this study than it did for the cadets at the USMA. Butler and McCauley (1987) identified other possible reasons for validity stability at the USMA, to include the use of common syllabi and tests for single courses taught by multiple instructors. Whether similar practices existed at civilian institutions beyond the required courses in the first few semesters, if at all, was not explored in this study.

Although the STEM groups in this study did not quite have the degree of validity stability found at the USMA, the amount of validity decay for the STEM majors was less than that for their non-STEM counterparts and the samples included in the Humphreys (1968) study, and this is an encouraging finding. The results suggest that validity decay is related to how observations are grouped. Validity decay was reduced by separating the STEM majors from the non-STEM majors, but within the non-STEM group were academic majors that were probably not as similar as those found within the STEM-Quantitative and STEM-Biological groups. Within the non-STEM group mean ACT scores and HSGPAs probably varied across majors, and the criterion probably varied from one non-STEM field to the next. While the STEM fields provided a good starting point, future analyses should be conducted by splitting apart the non-STEM majors.

Beyond the contributions to the research literature, the results of this study have applied contributions. There is national concern that success in the STEM fields has implications for the nation's economic success (e.g., National Governor's Association, 2007; National Science and Technology Council, 2013). The results of this study demonstrate that there are strong relationships between precollege predictors, ACT scores and HSGPA, with academic outcomes for STEM majors well beyond the first year of college. For the eight STEM subgroups in the hierarchical moderator analysis, all corrected validity coefficients for ACT Composite scores exceeded .40 (Table 5) in the eighth semester, and all corrected validity coefficients for HSGPA exceeded .55 (Table 10). Students entering college with high ACT scores and HSGPAs tend to earn higher course grades than their peers entering college with

lower ACT scores and HSGPAs not just in the first year but out to the end of the fourth year. This is especially true in the STEM fields.

Given the focus on the STEM fields, this study sends an important message to government officials, college admissions officers, high school counselors, teachers, parents and students. It is imperative that students intending to declare a STEM field as their major arrive at college prepared for the rigors they will face in undergraduate STEM programs. Students who do not arrive at college prepared may find it difficult to earn passing grades and to continue in a STEM program until graduation. This is not to say that students with low test scores and high school grades should be denied the opportunity to attempt studies in a STEM program. The point is that all parties with an interest in the STEM fields should have a realistic perspective on which students are more likely to succeed over four years in a STEM program given the students' precollege academic achievement. They should not expect entering students with low test scores and poor high school grades to be highly successful in STEM programs at the college level. College is not the time to play catch-up.

References

ACT. (2007). *The ACT technical manual*. Iowa City, IA: ACT.

ACT. (2009). *ACT interest inventory technical manual*. Iowa City, IA: ACT.

ACT. (2010b). *National collegiate retention and persistence to degree rates*. Iowa City, IA: ACT.

Allen, J., & Robbins, S. (2010). Effects of interest-major congruence, motivation, and academic performance on timely degree attainment. *Journal of Counseling Psychology, 57*(1), 23-35.

Allen, J., Robbins, S., Casillas, A., & Oh, I. (2008). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, 49, 647-664.

Austin, J. T., Humphreys, L. G., & Hulin, C. L. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology, 42*, 583-596.

Barrett, G. V., & Alexander, R. A. (1989). Rejoinder to Austin, Humphreys, and Hulin: Critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology, 42*, 597-612.

Barrett, G. V., Alexander, R. A., & Doverspike, D. (1992). The implications for personnel selection of apparent declines in predictive validities over time: A critique of Hulin, Henry, and Noon. *Personnel Psychology, 45*, 601-617.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38*, 41-56.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability: Their status 20 years later. *Psychological Science, 11,* 474–480.

Berry, C. M., & Sackett, P. R. (2009). Individual course choice result in underestimation of the validity of college admission systems. *Psychological Science, 20*, 822-830.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education, Praeger.

Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses* (College Board Research Report 2008-1, ETS RR-08-06). New York, NY: The College Board.

Burnham, P. S., & Hewitt, B. A. (1972). Quantitative factor scores as predictors of general academic promise. *Educational and Psychological Measurement, 32*, 403-410.

Butler, R. P., & McCauley, C. (1987). Extraordinary stability and ordinary predictability of academic success at the United States Military Academy. *Journal of Educational Psychology, 79*(1), 83-86.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action.* New York, NY: North-Holland.

Ceci, S. J. (Ed); & Williams, W. M. (Ed), (2007). *Why aren't more women in science: Top researchers debate the evidence*. Washington, DC: American Psychological Association

Chen, X., & Weko, T. (2009). *Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education* (NCES 2009-161). National Center for Education Statistics. Washington, DC: U. S. Department of Education.

Cronbach, L. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper & Row Publishers.

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

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

Goldman, R. D., & Hewitt, B. N. (1976). The Scholastic Aptitude Test "explains" why college men major in Science more often than college women. *Journal of Counseling Psychology, 23,* 50-54.

Goldman, R. D., Schmidt, D. E., Hewitt, B. N., & Fisher, R. (1974). Grading practices in different major fields. *American Education Research Journal, 11*(4), 343-357.

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*, 381-390.

Green, K. C. (1989). A profile of undergraduates in the sciences. *American Scientist, 77*, 475-480.

Green, M. (2007). *Science and Engineering Degrees: 1966-2004 (NSF 07-307)*. Arlington, VA: National Science Foundation.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum.

Hewitt, B. N., & Jacobs, R. (1978). Student perceptions of grading practices in different major fields. *Journal of Educational Measurement, 15,* 213-218.

Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika, 25,* 313-323.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *Journal of Educational Psychology, 59*, 375-380.

Humphreys, L. G. (1976). The phenomena are ubiquitous – but the investigator must look. *Journal of Educational Psychology, 68*(5), 521-521.

Humphreys, L. G., & Taber, T. (1973). Postdiction study of the GRE and eight semesters of college grades. *Journal of Educational Measurement, 10*, 179-184.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594-612.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education, Praeger.

Kimura, D. (2007). Underrepresentation or misrepresentation? In S. Ceci & W. Williams (Eds.), *Why aren't more women in science: Top researchers debate the evidence* (pp. 39-46). Washington, DC: American Psychological Association.

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.

Klitgaard, R. (1985). *Choosing elites*. New York, NY: Basic Books.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *The validity of the SAT for predicting first-year grade point average*. (College Board Research Report 2008-5). New York, NY: The College Board.

Kokkelenberg, E. C., & Sinha, E. (2010). Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students. *Economics of Education Review, 29,* 935-946.

Le, H., Casillas, A., Robbins, S.B., & Langley R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the student readiness inventory. *Educational and Psychological Measurement,* 65(3): 482-508.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods, 11*(4), 416-438.

Lubinski, D. S., & Benbow, C. P. (2007). Sex differences in personal attributes for the development of scientific expertise. In S. Ceci & W. Williams (Eds.), *Why aren't more women in science: Top researchers debate the evidence*, (pp. 79-100). Washington, DC: American Psychological Association.

Mattern, K. D., & Patterson, B. F. (2011a). *Validity of the SAT for predicting second-year grades: 2006 SAT validity sample*. (College Board Statistical Report 2011-1). New York: College Board.

Mattern, K. D., & Patterson, B. F. (2011b). *Validity of the SAT for predicting third-year grades: 2006 SAT validity sample*. (College Board Statistical Report 2011-3). New York, NY: College Board.

Mattern, K. D., & Patterson, B. F. (2011c). *Validity of the SAT for predicting fourth-year grades: 2006 SAT validity sample*. (College Board Statistical Report 2011-7). New York, NY: College Board.

Mauger, P. A., & Kolmodin, C. A. (1975). Long-term predictive validity of the Scholastic Aptitude Test. *Journal of Educational Psychology, 67*, 847-851.

National Center for Education Statistics (2002). Classification of Instructional Programs – 2000. (2002). *US Department of Education National Center for Education Statistics* (NCES 2002-165). Washington, DC: US Government Printing Office.

National Governors Association. (2007). *Building a science, technology, engineering and math agenda*. Washington, DC: National Governors Association.

National Science Board. (2010). *Science and engineering indicators 2010* (NSB 10-01). Arlington, VA: National Science Foundation.

National Science and Technology Council. (2013, May 31). *Federal science, technology, engineering, and mathematics (STEM) education 5-year strategic plan*. Retrieved July 15, 2013 from http://www.whitehouse.gov/sites/default/files/microsites/ostp/stem_stratplan_2013.pdf.

Nicholls, G. M., Wolfe, H., Besterfield-Sacre, M. Shuman, L. J., & Larpkiattaworn (2007). A method for identifying variables for predicting STEM enrollment. *Journal of Engineering Education, 96*, 33-43.

Noble, J., & Sawyer, R. (1987). Predicting grades in specific college freshman courses from ACT test scores and self-reported high school grades (ACT Research Report No. 87-20). Iowa City, IA: ACT.

Oh, E. (1976). *A study of instructor grading philosophies and practices in undergraduate courses at Western Michigan University*. (Doctoral dissertation). Retrieved from Dissertation and Thesis database. (UMI No. 302847348)

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review, 29*, 923-934.

Pennock-Roman, M. (1994). *College major and gender differences in the prediction of college grades* (College Board Report No. 94-2). New York, NY: The College Board.

Powers, D. E. (1982). Long-term predictive and construct validity of two predictors of law school performance. *Journal of Educational Psychology, 74*, 568-576.

Prather, J. E., & Smith, G. (1976). A study of the relationships between faculty characteristics, subject field, and course grading patterns. *Research in Higher Education, 5,* 351-363.

Prather, J. E., Smith, G., & Kodras, J. E. (1979). A longitudinal study of grades in 144 undergraduate courses. *Research in Higher Education, 10,* 11-24.

Price, J. (2010). The effect of instructor race and gender on student persistence in STEM fields. *Economics of Education Review, 29*, 901-910.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting grades: Sex, language, and ethnic groups*. (College Board Research Report 93-1). New York, NY: The College Board.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admission tests and post-secondary academic performance? *Psychological Bulletin, 135*(1), 1-22.

Schiel, J., & Noble, J. (1991). *Accuracy of self-reported course work and grade information of high school sophomores* (ACT Research Report 91-6). Iowa City, IA: ACT.

Snow, C. P. (1959). *The two cultures and the scientific revolution*. Cambridge, UK: Cambridge University Press.

Steele-Johnson, D., Osburn, H. G., & Pieper, K. F. (2000). A review and extension of current models of dynamic criteria. *International Journal of Selection and Assessment, 8*(3), 110-126.

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

Strenta, A. C., Elliot, R., Adair, R., Matier, M., & Scott, J. (1994). Choosing and leaving science in highly selective institutions. *Research in Higher Education, 35*(5), 513-547.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology, 85*(4), 710-718.

Thorndike, R. L. (1963). *The concepts of over- and underachievement*. New York, NY: Bureau of Publication, Teachers College, Columbia University.

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence 38*, 412-423.

White, P. E. (1992). *Women and minorities in science and engineering: An update* (NSF-92-303). Washington, DC: National Science Foundation.

Wilson, K. M. (1978). *Predicting the long-term performance in college of minority and nonminority students: A comparative analysis in two collegiate settings*. (RDR 77-78, No. 3 and RB-78-6). Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1980). The performance of minority students beyond the freshman year: Testing a "late bloomer" hypothesis in one Ohio state university setting. *Research in Higher Education 13*, 23-47.

Wilson, K. M. (1981). Analyzing the long-term performance of minority and nonminority students in: A tale of two studies. *Research in Higher Education 15*, 351-375.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year.* (College Board Research Report No. 83-2). New York, NY: The College Board.

Young, J. W. (1990a). Adjusting the cumulative GPA using Item response Theory. *Journal of Educational Measurement, 27,* 175-186.

Young, J. W. (1990b). Are validity coefficients understated due to correctable defects in the GPA? *Research in Higher Education, 31,* 319-325.

Young, J. W. (1993). Grade adjustment methods. *Review of Educational Research, 63*, 151-165.

Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis*. (College Board Research Report No. 2001-6). New York, NY: The College Board.

Zwick, R. (2006). Higher education admission testing. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 647-679). Westport, CT: American Council on Education, Praeger.

Footnotes

1 Although the enrolled cohorts used were from 2000 to 2005, examinees for the reference population ranged from 1999 to 2005 because most students take the ACT at the end of their junior year or beginning of their senior year, which would have been 1999 for the first cohort year.

Appendix

Table 1

Previous Research on Validity Decay/Stability across Four Years

Study/Institution(s)	*N*	Measure	Semester 1	2	3	4	5	6	7	8	Change	Percent Change
Humphreys (1968)	1,600[a]	ACTC	.38	.30	.24	.26	.24	.25	.22	.17	-.20	52.6
University of Illinois		ACTE	.35	.26	.23	.24	.24	.22	.22	.16	-.19	-54.3
		ACTM	.28	.19	.17	.17	.15	.16	.16	.12	-.16	-57.1
		ACTSS	.28	.24	.19	.20	.21	.23	.17	.15	-.13	-46.4
		ACTNS	.31	.26	.18	.20	.18	.20	.16	.13	-.18	-58.1
		HSR	.39	.34	.28	.27	.24	.26	.24	.22	-.17	-43.6
Humphreys (1968)	Max. N	ACTC	.48	.38	.28	.26	.24	.24	.20	.16	-.32	-66.3
University of Illinois	in each	ACTE	.40	.33	.23	.21	.19	.18	.19	.15	-.25	-62.3
	semester	ACTM	.40	.29	.22	.19	.17	.16	.14	.11	-.28	-71.7
		ACTSS	.37	.31	.24	.22	.20	.22	.16	.13	-.26	-64.7
		ACTNS	.36	.28	.21	.20	.19	.18	.16	.12	-.24	-66.8
		HSR	.51	.42	.32	.30	.27	.26	.24	.22	-.30	-57.9
Humphreys (1968)[b]	1,600	ACTC	.47	.40	.32	.31	.28	.29	.25	.21	-.26	-54.9
University of Illinois		ACTE	.40	.35	.27	.25	.22	.22	.24	.20	-.20	-50.0
		ACTM	.40	.30	.25	.23	.20	.20	.18	.15	-.25	-63.1
		ACTSS	.37	.33	.27	.26	.24	.27	.19	.17	-.20	-53.4
		ACTNS	.36	.30	.24	.23	.22	.23	.20	.16	-.20	-56.0
		HSR	.51	.45	.37	.36	.31	.32	.30	.28	-.23	-44.8
Humphreys & Taber (1973)	1,510	GRE-V	.35	.31	.26	.27	.25	.22	.21	.16	-.19	-54.3
University of Illinois	1,510	GRE-Q	.35	.33	.31	.29	.28	.20	.17	.15	-.20	-57.1
	987	GRE-Ad.	.36	.39	.38	.38	.34	.33	.31	.23	-.13	-36.1
Wilson (1978, 1980)	530[a]	SAT-V	.42	.42	.46	.48	.47	.40	.38	.37	-.05	-11.9
Anonymous university		SAT-M	.42	.40	.47	.45	.49	.37	.30	.32	-.10	-23.8
		HSR	.51	.45	.45	.41	.45	.40	.37	.36	-.15	-29.4

Table 1 (cont.)

Previous Research on Validity Decay/Stability across Four Years

			Semester									
Study/Institution(s)	*N*	Measure	1	2	3	4	5	6	7	8	Change	Percent Change
Wilson (1978, 1981) [c]	950[a]	CB Ac. Av.		.51	.45	.40	.42	.39	.40	.34	-.11[d]	-24.4
Selective liberal arts college		SAT-V		.42	.40	.34	.39	.35	.35	.33	-.07[d]	-17.5
		SAT-M		.38	.36	.29	.32	.28	.28	.20	-.16[d]	-42.1
		HSR		.36	.29	.28	.27	.25	.19	.20	-.09[d]	-31.0
Butler & McCauley (1987)	559	SAT-V	.30	.16	.26	.27	.28	.25	.31	.25	-.05	-16.7
USAFA		SAT-M	.41	.36	.39	.36	.36	.28	.27	.25	-.16	-39.0
		HSR	.39	.39	.47	.41	.44	.41	.39	.36	-.03	-7.7
Butler & McCauley (1987)[e]	618	SAT-V		.30		.30		.30		.30	.00	0.0
USMA, 1982 class		SAT-M		.41		.43		.43		.42	+.01	+2.4
		HSR		.47		.51		.51		.51	+.04	+8.5
Butler & McCauley (1987)[e]	631	SAT-V		.36		.32		.32		.32	-.04	-11.1
USMA, 1983 class		SAT-M		.35		.40		.39		.39	+.04	+11.4
		HSR		.38		.41		.41		.41	+.03	+7.9
Elliot & Strenta (1988)[e]	927	SAT-V		.35		.32		.31		.32	-.03	-8.6
Dartmouth College		SAT-M		.39		.31		.25		.24	-.15	-38.5
		HSR		.42		.34		.35		.31	-.11	-26.2
		SAT-V-A		.36		.34		.33		.35	-.01	-2.8
		SAT-M-A		.46		.40		.36		.37	-.09	-19.6
		HSR-A		.44		.39		.39		.36	-.03	-6.8

Table 1 (cont.)

Previous Research on Validity Decay/Stability across Four Years

			Semester									
Study/Institution(s)	*N*	Measure	1	2	3	4	5	6	7	8	Change	Percent Change
Kobrin et al. (2008)[f]	151,316[f]	SAT-CR		.29[f]		.27[g]		.23[h]		.20[i]	-.09	-31.0
110 institutions		SAT-M		.26[f]		.23[g]		.18[h]		.15[i]	-.11	-42.3
Mattern & Patterson (2011a)[g]	75,208[g]	SAT-W		.33[f]		.31[g]		.27[h]		.24[i]	-.09	-27.3
66 institutions		SAT-CR/M		.32[f]		.29[g]		.24[h]		.21[i]	-.11	-34.4
Mattern & Patterson (2011b)[h]	63,736[h]	SAT-CR/M/W		.35[f]		.32[g]		.28[h]		.24[i]	-.11	-31.4
60 institutions		HSGPA		.36[f]		.32[g]		.29[h]		.27[i]	-.09	-25.0
Mattern & Patterson (2011c)[i]	56,939[i]	SAT-CR-C		.48[f]		.45[g]		.40[h]		.35[i]	-.13	-.27.1
55 institutions		SAT-M-C		.47[f]		.44[g]		.38[h]		.33[i]	-.14	-29.8
		SAT-W-C		.51[f]		.49[g]		.43[h]		.39[i]	-.12	-23.5
		SAT-CR/M-C		.51[f]		.48[g]		.42[h]		.37[i]	-.14	-27.5
		SAT-CR/M/W-C		.53[f]		.50[g]		.45[h]		.40[i]	-.13	-24.5
		HSGPA-C		.54[f]		.51[g]		.46[h]		.43[i]	-.11	-20.4

Note: [a] This is an approximation, as the values varied slightly across measures and semesters; [b] Corrected for range restriction; [c] Correlation of predictor with Year 1 GPA; [d] Change from third semester to eighth semester; [e] Correlations with independently calculated annual GPAs instead of independently calculated semester GPAs; [f] Correlations with independently calculated first-year GPAs, Kobrin et al. (2008); [g] Correlations with independently calculated second-year GPAs, Mattern & Patterson (2011a); [h] Correlations with independently calculated third-year GPAs, Mattern & Patterson (2011b); [i] Correlations with independently calculated fourth-year GPAs, Mattern & Patterson (2011c); USAFA = United States Air Force Academy; USMA = United States Military Academy; ACTC = ACT Composite; ACTE = ACT English; ACTM = ACT Mathematics; ACTSS = ACT Social Studies; ACTNS = ACT Natural Science; HSR = High School Rank; GRE-V = Graduate Record Exam, Verbal; GRE-Q = Graduate Record Exam, Quantitative; GRE-Ad. = Graduate Record Exam, Advanced Tests; CB Ac. Av. = College Board Achievement Tests, Average; SAT-V = SAT Verbal; SAT-M = SAT Mathematics; SAT-V-A = SAT Verbal Adjusted; SAT-M-A = SAT Mathematics Adjusted; HSR-A = High School Rank Adjusted; SAT-CR = SAT Critical Reading; SAT-W = SAT Writing; SAT-CR/M = SAT Critical Reading and Mathematics, multiple correlation; SAT-CR/M/W = SAT Critical Reading, Mathematics, and Writing, multiple correlation; HSGPA = high school grade point average; SAT-CR-C = SAT-Critical Reading, corrected for range restriction; SAT-M-C = SAT-Mathematics, corrected for range restriction; SAT-W-C = SAT-Writing, corrected for range restriction; SAT-CR/M-C = SAT Critical Reading and Mathematics, multiple correlation, corrected for range restriction; SAT-CR/M/W-C = SAT Critical Reading, Mathematics, and Writing, multiple correlation, corrected for range restriction; HSGPA-C = high school grade point average, corrected for range restriction.

Table 2

Typical range of ACT Composite Scores and Class Ranks by Institution Admission Selectivity

Institution Selectivity Level	ACT Composite Scores Middle 50%	Definition
1. Highly Selective	25—30	Majority admitted from top 10% of high school class
2. Selective	21—26	Majority admitted from top 25% of high school class
3. Traditional	18—24	Majority admitted from top 50% of high school class
4. Liberal	17—22	Majority admitted from bottom 50% of high school class
5. Open	16—21	Generally open to all with high school diploma or equivalent

Note. ACT Composite score scale ranges from 1 to 36. Adapted from *National Collegiate Retention and Persistence to Degree Rates* (ACT, 2010b). Means and SDs calculated from ACT 2010-2011 examinees.

Table 3

Reference Populations' Means, Standard Deviations, and Correlations between Precollege Academic Predictors, ACT National Data, 1999-2006

	Measure	*N*	*Mean*	*SD*	ACTC	ACTE	ACTM	ACTR	ACTS	HSGPA
Overall	ACTC	7,990,217	20.9	4.8	1.00					
	ACTE	7,990,217	20.4	5.8	.91	1.00				
	ACTM	7,990,217	20.7	5.0	.87	.72	1.00			
	ACTR	7,990,217	21.3	6.1	.90	.79	.65	1.00		
	ACTS	7,990,217	20.9	4.6	.89	.73	.76	.72	1.00	
	HSGPA	6,625,660	3.21	0.61	.58	.54	.56	.48	.50	1.00
Male	ACTC	3,452,926	21.1	5.0	1.00					
	ACTE	3,452,926	19.9	5.8	.92	1.00				
	ACTM	3,452,926	21.3	5.3	.88	.74	1.00			
	ACTR	3,452,926	21.0	6.1	.90	.79	.66	1.00		
	ACTS	3,452,926	21.4	4.9	.90	.75	.77	.75	1.00	
	HSGPA	2,792,352	3.12	0.62	.59	.53	.59	.47	.52	1.00
Female	ACTC	4,448,885	20.9	4.7	1.00					
	ACTE	4,448,885	20.8	5.7	.92	1.00				
	ACTM	4,448,885	20.2	4.8	.86	.73	1.00			
	ACTR	4,448,885	21.5	6.0	.90	.79	.65	1.00		
	ACTS	4,448,885	20.5	4.3	.88	.74	.75	.72	1.00	
	HSGPA	3,783,691	3.28	0.58	.59	.54	.58	.48	.52	1.00

Note: Overall figures include examinees who did not identify their gender. All correlations are significant at $\underline{p} < .0001$. ACTC = ACT Composite; ACTE = ACT English; ACTM = ACT Mathematics; ACTR = ACT Reading; ACTS = ACT Science; HSGPA = High school grade point average.

Table 4

Means (Standard Deviations) for ACT Scores and HSGPA, Overall and by Subgroups

SMC	Gender	Admission Selectivity	*k*	*N*	ACTC	ACTE	ACTM	ACTR	ACTS	HSGPA
All	All	All	26	62,212	23.6 (4.2)	23.8 (5.0)	23.1 (4.8)	24.1 (5.5)	23.0 (4.1)	3.58 (0.43)
STEM-Quant	Female	More	10	1,386	27.0 (3.9)	27.0 (4.7)	27.7 (4.0)	26.9 (5.4)	25.7 (4.1)	3.85 (0.22)
STEM-Quant	Male	More	10	4,231	26.9 (3.9)	26.0 (4.7)	28.3 (3.9)	26.3 (5.4)	26.6 (4.3)	3.76 (0.29)
STEM-Bio	Female	More	10	1,621	25.9 (3.5)	26.3 (4.4)	25.7 (3.9)	26.4 (5.0)	24.5 (3.7)	3.82 (0.24)
STEM-Bio	Male	More	10	1,015	26.4 (3.6)	26.0 (4.5)	27.0 (3.9)	26.2 (5.0)	25.9 (4.0)	3.77 (0.27)
Non-STEM	Female	More	10	23,127	23.8 (3.7)	24.6 (4.6)	22.8 (4.1)	24.6 (5.2)	22.6 (3.6)	3.65 (0.34)
Non-STEM	Male	More	10	12,922	23.9 (3.9)	23.7 (4.7)	23.8 (4.3)	24.2 (5.4)	23.6 (3.9)	3.54 (0.40)
STEM-Quant	Female	Less	16	174	24.1 (3.9)	24.1 (5.1)	25.1 (4.2)	23.4 (5.3)	23.5 (3.6)	3.74 (0.33)
STEM-Quant	Male	Less	16	672	24.3 (3.8)	22.7 (4.5)	25.6 (4.1)	23.7 (5.3)	24.7 (4.1)	3.58 (0.39)
STEM-Bio	Female	Less	16	409	23.3 (3.9)	23.7 (4.9)	22.2 (4.2)	24.0 (5.3)	22.9 (3.8)	3.68 (0.35)
STEM-Bio	Male	Less	16	194	23.7 (4.0)	23.1 (4.6)	23.6 (4.3)	23.7 (5.8)	23.9 (3.7)	3.65 (0.40)
Non-STEM	Female	Less	16	10,350	21.6 (3.7)	22.0 (4.6)	20.5 (3.9)	22.4 (5.2)	21.2 (3.5)	3.48 (0.44)
Non-STEM	Male	Less	16	6,021	21.5 (3.7)	20.7 (4.6)	21.3 (4.2)	21.8 (5.2)	21.8 (3.7)	3.31 (0.48)

Note. ACTC=ACT Composite; ACTE=ACT English; ACTM=ACT Mathematics; ACTR=ACT Reading; ACTS=ACT Science; HSGPA=high school grade point average; *k*=number of institutional studies; STEM=science, technology, engineering, and, mathematics; Quant=quantitative; Bio=biological.

Table 5

Uncorrected and Corrected Correlations, ACT-Composite Scores and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.44	.42	.45	.39	.39	.35	.39	.38	-.05	-12.1
			STEM-Biological	10	1,621	.37	.34	.31	.30	.30	.28	.27	.26	-.10	-28.1
			Non-STEM	10	23,127	.42	.41	.41	.37	.32	.27	.27	.22	-.20	-46.7
		Less	STEM-Quantitative	16	174	.34	.41	.45	.41	.48	.39	.46	.45	.11	31.9
			STEM-Biological	16	409	.48	.44	.48	.40	.41	.33	.33	.35	-.13	-26.8
			Non-STEM	16	10,350	.43	.44	.43	.40	.37	.33	.29	.26	-.17	-40.1
	Male	More	STEM-Quantitative	10	4,231	.39	.35	.36	.31	.33	.28	.27	.26	-.14	-34.3
			STEM-Biological	10	1,015	.37	.36	.27	.28	.29	.23	.23	.26	-.11	-30.7
			Non-STEM	10	12,922	.36	.36	.37	.32	.28	.24	.25	.19	-.16	-45.3
		Less	STEM-Quantitative	16	672	.33	.35	.42	.33	.31	.29	.26	.29	-.04	-11.6
			STEM-Biological	16	194	.29	.29	.27	.16	.27	.13	.29	.28	-.01	-3.8
			Non-STEM	16	6,021	.38	.39	.39	.35	.33	.28	.27	.23	-.15	-38.5
Corrected	Female	More	STEM-Quantitative	10	1,386	.61	.60	.63	.58	.59	.53	.55	.55	-.07	-10.7
			STEM-Biological	10	1,621	.57	.53	.50	.48	.50	.48	.42	.42	-.15	-26.5
			Non-STEM	10	23,127	.64	.63	.63	.57	.51	.43	.42	.35	-.29	-45.2
		Less	STEM-Quantitative	16	174	.58	.61	.65	.60	.71	.59	.62	.62	.04	7.2
			STEM-Biological	16	409	.68	.63	.68	.61	.59	.47	.46	.48	-.20	-29.3
			Non-STEM	16	10,350	.67	.67	.65	.62	.56	.51	.45	.41	-.26	-39.2
	Male	More	STEM-Quantitative	10	4,231	.62	.56	.58	.53	.55	.47	.42	.41	-.21	-33.3
			STEM-Biological	10	1,015	.59	.58	.45	.45	.46	.39	.39	.42	-.17	-29.1
			Non-STEM	10	12,922	.58	.58	.59	.52	.46	.40	.40	.32	-.26	-45.6
		Less	STEM-Quantitative	16	672	.55	.57	.66	.53	.50	.47	.40	.43	-.12	-21.7
			STEM-Biological	16	194	.45	.46	.43	.26	.43	.21	.45	.45	.01	1.5
			Non-STEM	16	6,021	.63	.65	.64	.58	.54	.47	.45	.39	-.24	-37.9

Note: *k*=number of studies; STEM=science, technology, engineering, and mathematics.

Table 6

Uncorrected and Corrected Correlations, ACT-English Scores and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.39	.39	.40	.34	.34	.30	.34	.35	-.04	-10.9
			STEM-Biological	10	1,621	.32	.30	.27	.28	.28	.25	.23	.24	-.08	-24.5
			Non-STEM	10	23,127	.38	.37	.37	.34	.29	.25	.25	.21	-.17	-45.1
		Less	STEM-Quantitative	16	174	.35	.44	.38	.32	.45	.37	.46	.42	.07	19.9
			STEM-Biological	16	409	.44	.42	.44	.36	.41	.33	.30	.32	-.12	-27.0
			Non-STEM	16	10,350	.40	.40	.38	.36	.34	.30	.27	.24	-.16	-39.7
	Male	More	STEM-Quantitative	10	4,231	.36	.34	.35	.30	.30	.26	.24	.24	-.12	-33.1
			STEM-Biological	10	1,015	.34	.32	.26	.26	.26	.22	.21	.23	-.11	-32.4
			Non-STEM	10	12,922	.31	.32	.34	.29	.25	.21	.22	.18	-.14	-43.7
		Less	STEM-Quantitative	16	672	.34	.33	.39	.31	.28	.28	.23	.26	-.07	-21.2
			STEM-Biological	16	194	.31	.30	.29	.22	.27	.16	.29	.32	.02	5.6
			Non-STEM	16	6,021	.35	.36	.36	.31	.30	.25	.24	.21	-.14	-40.1
Corrected	Female	More	STEM-Quantitative	10	1,386	.58	.59	.59	.53	.55	.48	.52	.53	-.06	-9.6
			STEM-Biological	10	1,621	.53	.50	.45	.45	.48	.43	.38	.40	-.12	-23.4
			Non-STEM	10	23,127	.60	.58	.59	.53	.47	.41	.39	.34	-.26	-43.8
		Less	STEM-Quantitative	16	174	.55	.61	.50	.45	.60	.53	.59	.57	.01	1.9
			STEM-Biological	16	409	.64	.60	.62	.55	.59	.47	.41	.43	-.20	-31.9
			Non-STEM	16	10,350	.62	.63	.60	.57	.53	.48	.42	.38	-.24	-39.1
	Male	More	STEM-Quantitative	10	4,231	.57	.55	.56	.51	.50	.44	.38	.38	-.19	-33.3
			STEM-Biological	10	1,015	.54	.51	.42	.41	.41	.35	.33	.37	-.17	-31.4
			Non-STEM	10	12,922	.52	.52	.55	.47	.42	.35	.36	.29	-.23	-44.8
		Less	STEM-Quantitative	16	672	.55	.53	.61	.49	.44	.45	.34	.39	-.16	-29.3
			STEM-Biological	16	194	.49	.50	.44	.35	.45	.31	.47	.53	.05	9.3
			Non-STEM	16	6,021	.58	.59	.59	.52	.49	.42	.39	.35	-.23	-40.0

Note: *k*=number of studies; STEM=science, technology, engineering, and mathematics.

Table 7

Uncorrected and Corrected Correlations, ACT-Mathematics Scores and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.42	.40	.42	.40	.38	.33	.38	.36	-.06	-13.9
			STEM-Biological	10	1,621	.40	.35	.32	.30	.30	.29	.28	.26	-.14	-35.0
			Non-STEM	10	23,127	.40	.38	.37	.34	.30	.25	.25	.21	-.19	-47.4
		Less	STEM-Quantitative	16	174	.39	.46	.48	.44	.46	.42	.44	.44	.05	13.0
			STEM-Biological	16	409	.44	.42	.49	.38	.39	.28	.33	.30	-.13	-30.6
			Non-STEM	16	10,350	.39	.40	.39	.36	.33	.31	.27	.24	-.15	-38.2
	Male	More	STEM-Quantitative	10	4,231	.40	.32	.35	.31	.33	.28	.28	.26	-.14	-34.2
			STEM-Biological	10	1,015	.38	.38	.30	.30	.32	.25	.23	.27	-.11	-29.1
			Non-STEM	10	12,922	.34	.35	.34	.31	.27	.24	.23	.19	-.16	-46.0
		Less	STEM-Quantitative	16	672	.29	.34	.41	.31	.30	.25	.29	.27	-.02	-5.7
			STEM-Biological	16	194	.28	.33	.31	.14	.25	.15	.28	.24	-.04	-13.1
			Non-STEM	16	6,021	.34	.35	.35	.33	.30	.25	.24	.22	-.13	-36.7
Corrected	Female	More	STEM-Quantitative	10	1,386	.60	.59	.61	.61	.60	.52	.56	.54	-.07	-11.2
			STEM-Biological	10	1,621	.60	.53	.49	.46	.49	.47	.42	.40	-.20	-33.0
			Non-STEM	10	23,127	.59	.57	.55	.51	.45	.39	.37	.31	-.28	-47.0
		Less	STEM-Quantitative	16	174	.57	.65	.64	.60	.64	.59	.56	.58	.01	2.1
			STEM-Biological	16	409	.63	.61	.70	.56	.55	.41	.44	.43	-.20	-31.9
			Non-STEM	16	10,350	.59	.61	.59	.55	.50	.47	.41	.36	-.23	-38.9
	Male	More	STEM-Quantitative	10	4,231	.67	.57	.60	.57	.61	.53	.49	.46	-.21	-30.7
			STEM-Biological	10	1,015	.63	.63	.51	.50	.53	.42	.40	.45	-.17	-27.3
			Non-STEM	10	12,922	.56	.56	.56	.50	.44	.39	.36	.30	-.26	-46.6
		Less	STEM-Quantitative	16	672	.50	.57	.65	.53	.51	.43	.45	.42	-.08	-16.1
			STEM-Biological	16	194	.46	.53	.48	.26	.43	.35	.46	.40	-.06	-12.6
			Non-STEM	16	6,021	.57	.58	.57	.54	.48	.42	.40	.36	-.21	-36.9

Note: *k*=number of studies; STEM=science, technology, engineering, and mathematics.

Table 8

Uncorrected and Corrected Correlations, ACT-Reading Scores and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.33	.32	.36	.30	.30	.28	.30	.30	-.03	-9.1
			STEM-Biological	10	1,621	.24	.24	.24	.23	.22	.21	.19	.19	-.05	-21.2
			Non-STEM	10	23,127	.32	.31	.32	.29	.25	.20	.21	.16	-.15	-47.9
		Less	STEM-Quantitative	16	174	.24	.33	.37	.34	.41	.34	.40	.41	.17	72.3
			STEM-Biological	16	409	.38	.31	.35	.31	.31	.26	.25	.29	-.08	-22.1
			Non-STEM	16	10,350	.34	.33	.33	.30	.28	.25	.22	.19	-.15	-43.7
	Male	More	STEM-Quantitative	10	4,231	.30	.27	.28	.23	.25	.20	.19	.18	-.12	-39.0
			STEM-Biological	10	1,015	.23	.25	.17	.20	.19	.16	.16	.16	-.07	-32.3
			Non-STEM	10	12,922	.27	.27	.28	.25	.21	.18	.19	.14	-.13	-46.7
		Less	STEM-Quantitative	16	672	.24	.29	.32	.25	.24	.24	.19	.24	-.01	-2.9
			STEM-Biological	16	194	.23	.18	.18	.10	.20	.06	.20	.20	-.03	-13.2
			Non-STEM	16	6,021	.28	.30	.30	.26	.25	.21	.20	.17	-.12	-40.9
Corrected	Female	More	STEM-Quantitative	10	1,386	.47	.45	.50	.44	.44	.42	.42	.42	-.04	-9.2
			STEM-Biological	10	1,621	.38	.37	.37	.35	.36	.35	.29	.30	-.08	-21.6
			Non-STEM	10	23,127	.48	.47	.49	.43	.38	.31	.31	.25	-.23	-48.4
		Less	STEM-Quantitative	16	174	.41	.47	.49	.48	.56	.48	.54	.55	.14	33.6
			STEM-Biological	16	409	.55	.46	.50	.49	.45	.37	.37	.42	-.13	-23.0
			Non-STEM	16	10,350	.51	.50	.50	.46	.42	.37	.33	.28	-.23	-44.4
	Male	More	STEM-Quantitative	10	4,231	.45	.42	.42	.36	.39	.32	.28	.27	-.18	-39.5
			STEM-Biological	10	1,015	.38	.40	.29	.32	.30	.26	.26	.25	-.12	-32.8
			Non-STEM	10	12,922	.42	.42	.43	.37	.33	.28	.29	.22	-.21	-48.6
		Less	STEM-Quantitative	16	672	.40	.45	.48	.39	.38	.38	.27	.32	-.07	-18.0
			STEM-Biological	16	194	.31	.26	.23	.14	.24	.10	.29	.28	-.03	-9.4
			Non-STEM	16	6,021	.46	.48	.47	.42	.40	.34	.32	.26	-.19	-42.5

Note: *k*=number of studies; STEM=science, technology, engineering, and mathematics.

Table 9

Uncorrected and Corrected Correlations, ACT-Science Scores and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.36	.34	.36	.32	.31	.28	.31	.31	-.05	-13.7
			STEM-Biological	10	1,621	.26	.22	.20	.20	.20	.19	.20	.18	-.08	-29.9
			Non-STEM	10	23,127	.33	.33	.32	.30	.26	.22	.22	.18	-.16	-46.7
		Less	STEM-Quantitative	16	174	.28	.25	.33	.27	.33	.19	.27	.18	-.10	-35.4
			STEM-Biological	16	409	.39	.38	.38	.33	.31	.24	.26	.27	-.12	-30.0
			Non-STEM	16	10,350	.36	.36	.35	.33	.30	.28	.24	.22	-.14	-38.1
	Male	More	STEM-Quantitative	10	4,231	.31	.27	.28	.25	.26	.22	.22	.21	-.10	-32.6
			STEM-Biological	10	1,015	.28	.25	.18	.17	.20	.15	.18	.21	-.07	-26.6
			Non-STEM	10	12,922	.28	.28	.28	.26	.22	.20	.20	.15	-.13	-45.0
		Less	STEM-Quantitative	16	672	.24	.24	.31	.25	.22	.20	.19	.22	-.02	-8.2
			STEM-Biological	16	194	.14	.14	.13	.06	.16	.05	.17	.18	.04	26.3
			Non-STEM	16	6,021	.31	.32	.31	.29	.27	.23	.22	.19	-.12	-38.1
Corrected	Female	More	STEM-Quantitative	10	1,386	.47	.45	.46	.43	.45	.39	.41	.41	-.07	-13.9
			STEM-Biological	10	1,621	.40	.35	.31	.31	.32	.31	.30	.28	-.12	-29.3
			Non-STEM	10	23,127	.54	.53	.53	.49	.43	.37	.36	.29	-.25	-45.8
		Less	STEM-Quantitative	16	174	.48	.42	.51	.43	.53	.27	.34	.29	-.19	-38.7
			STEM-Biological	16	409	.57	.56	.58	.52	.45	.34	.37	.38	-.20	-34.7
			Non-STEM	16	10,350	.59	.60	.59	.55	.50	.47	.40	.37	-.22	-37.8
	Male	More	STEM-Quantitative	10	4,231	.47	.43	.43	.41	.43	.35	.34	.32	-.16	-33.5
			STEM-Biological	10	1,015	.46	.41	.29	.27	.33	.24	.29	.34	-.12	-26.0
			Non-STEM	10	12,922	.50	.50	.51	.45	.39	.36	.35	.27	-.23	-45.4
		Less	STEM-Quantitative	16	672	.40	.37	.49	.39	.36	.32	.27	.32	-.08	-19.8
			STEM-Biological	16	194	.32	.33	.22	.12	.36	.11	.37	.34	.02	4.7
			Non-STEM	16	6,021	.58	.60	.57	.54	.51	.43	.41	.36	-.22	-37.4

Note: *k*=number of studies; STEM=science, technology, engineering, and mathematics.

Table 10

Uncorrected and Corrected Correlations, HSGPA and Semester GPA

Correlation	Gender	Admission Selectivity	Student Major Category	*k*	*N*	Semesters 1	2	3	4	5	6	7	8	Change 1-8	Percent Change
Uncorrected	Female	More	STEM-Quantitative	10	1,386	.34	.32	.32	.31	.27	.25	.30	.32	-.02	-6.1
			STEM-Biological	10	1,621	.39	.37	.32	.29	.28	.25	.24	.24	-.15	-39.0
			Non-STEM	10	23,127	.38	.38	.36	.33	.30	.27	.27	.24	-.14	-37.1
		Less	STEM-Quantitative	16	174	.42	.42	.33	.37	.37	.39	.35	.36	-.06	-14.5
			STEM-Biological	16	409	.52	.52	.53	.40	.38	.34	.40	.45	-.07	-13.7
			Non-STEM	16	10,350	.44	.45	.42	.41	.39	.36	.35	.31	-.13	-29.2
	Male	More	STEM-Quantitative	10	4,231	.34	.33	.29	.27	.26	.27	.25	.25	-.10	-27.8
			STEM-Biological	10	1,015	.37	.38	.29	.27	.34	.24	.24	.26	-.11	-30.4
			Non-STEM	10	12,922	.37	.37	.36	.35	.31	.28	.27	.23	-.14	-38.1
		Less	STEM-Quantitative	16	672	.39	.43	.43	.38	.32	.34	.33	.37	-.01	-3.7
			STEM-Biological	16	194	.41	.45	.41	.37	.48	.27	.47	.35	-.06	-14.5
			Non-STEM	16	6,021	.41	.42	.40	.37	.36	.32	.31	.27	-.14	-34.9
Corrected	Female	More	STEM-Quantitative	10	1,386	.83	.81	.81	.82	.79	.75	.79	.81	-.02	-2.3
			STEM-Biological	10	1,621	.85	.83	.79	.75	.75	.71	.67	.67	-.19	-21.9
			Non-STEM	10	23,127	.73	.74	.71	.67	.64	.58	.56	.51	-.22	-29.9
		Less	STEM-Quantitative	16	174	.71	.74	.68	.62	.63	.70	.62	.60	-.11	-15.9
			STEM-Biological	16	409	.85	.86	.85	.77	.72	.66	.70	.76	-.09	-10.2
			Non-STEM	16	10,350	.70	.71	.68	.66	.62	.59	.56	.52	-.19	-26.5
	Male	More	STEM-Quantitative	10	4,231	.80	.78	.73	.72	.72	.72	.65	.65	-.15	-19.1
			STEM-Biological	10	1,015	.82	.83	.74	.69	.79	.66	.67	.68	-.14	-16.7
			Non-STEM	10	12,922	.72	.72	.71	.67	.62	.57	.54	.47	-.24	-34.0
		Less	STEM-Quantitative	16	672	.74	.78	.78	.72	.63	.65	.60	.65	-.08	-11.4
			STEM-Biological	16	194	.71	.73	.67	.67	.71	.49	.68	.58	-.13	-17.7
			Non-STEM	16	6,021	.68	.69	.66	.63	.60	.55	.51	.46	-.23	-33.2

Note: k=number of studies; STEM=science, technology, engineering, and mathematics.