



Iowa Testing Programs

Establishing Empirical Links between High School Assessments and College Outcomes
An Essential Requirement for College Readiness Interpretations
ITP Research Series

**Anthony David Fina
Catherine J. Welch
Stephen B. Dunbar**

ITP Research Series

2017.3

Abstract

As states evaluate whether they should continue with their current assessment program or adopt next-generation college readiness assessments, it is important to ascertain the degree to which current high school assessments can be used for college readiness interpretations. In this study, we examined the ability of a state assessment to serve as an indicator of college readiness. Empirical evidence is presented summarizing relationships between performance on the standards-based high school assessment and performance in college. Benchmarks were set on the Reading, Mathematics, and Science tests by linking assessment scores directly to grades in college courses. The accuracy of the benchmarks was similar to that of a traditional college admission test. Students who met the college readiness benchmarks earned higher grades in general education college courses and had higher first-year college grade point averages. Implications for states and other stakeholders are discussed.

Establishing Empirical Links between High School Assessments and College Outcomes: An Essential Requirement for College Readiness Interpretations

The Every Student Succeeds Act (ESSA) of 2015 requires that states establish challenging college- and career-ready (CCR) standards to ensure that all students graduate from high school ready for college or a career. Many states have adopted the Common Core State Standards (CCSS), or a modified version of the CCSS, to meet this requirement from the perspective of curriculum and instruction. The assessments used to measure the CCR standards, however, vary more than perhaps was expected by the developers of the CCSS. Although some 20 states in the 2016-17 school year used some version of the assessments developed by the Smarter Balanced Assessment Consortium (SBAC) or the Partnership for the Assessment of Readiness for College and Careers (PARCC), 30 others used their own assessments aligned with the state's standards (Viadero, 2017). An ultimate goal for the validation of all of these assessments is to bolster the evidence-based design arguments related to content alignment with empirical evidence that supports their validity with respect to the CCR construct and actual performance in postsecondary education. The purpose of this paper is to execute a methodology to articulate the development of such empirical evidence and to evaluate the extent to which a standards-based assessment used in a statewide testing program can serve as a basis for empirical benchmarks of readiness in three general areas of postsecondary education.

Camara (2013) outlines the types of evidence needed to support college readiness interpretations. While a comprehensive validity argument will consist of many sources of evidence supporting desired score interpretations (Kane, 2013), Camara argues that empirical evidence of the relationship between test scores and criterion measures of college success is necessary to support college readiness interpretations. General studies of the predictive validity

of admissions tests are familiar to institutional researchers. In recent years, such studies have been expanded to include predictive relationships between standards-based statewide assessments taken by students in high school and success in college. The present study seeks to augment the findings of this latter research with methods and results that focus on empirical aspects of a specific CCR inference, that is, to what extent does an assessment support an estimate of the probability that a given student will receive, for example, a C or better in specific postsecondary courses. This probability may be the most tangible, quantitative definition of what it means to say a student is able to perform at a level implied by the CCR standards. Although it is widely recognized that definitions of success after high school span a wide range of dimensions, both cognitive and non-cognitive, this quantitative definition can be argued to be a necessary component of a comprehensive program of validation research for any assessment that advances claims related to CCR standards.

Purpose of the Study

This study examines the extent to which a standards-based high school achievement test can be used as an indicator of readiness for postsecondary coursework. The analysis and results are framed by two objectives:

1. To establish empirical evidence of the relationship between a high school assessment and performance in postsecondary coursework.
2. To develop and evaluate readiness benchmarks on the high school assessment with respect to specific levels of performance.

The first objective serves to document the general strength of association between test scores and postsecondary performance as measured by first-year grade point average (GPA) to replicate

previous findings from the literature and establish a basis for readiness interpretations. The second objective leverages the bivariate relationship between test scores and course grades to set benchmarks on the state assessment, that is, cut-scores that would provide stakeholders with a basis for judging the probability a student would receive letter grades of C or B and higher in courses to which the standards-based test is related. This type of evidence is necessary to support postsecondary readiness interpretations from standards-based state assessments (Camara, 2013; Porter & Polikoff, 2012), required by professional standards (AERA, APA, & NCME, 2014), and is key component of peer review processes for federal accountability under the ESSA.

Literature Review

Traditionally, college admissions tests have been considered the primary indicators of college readiness for high school students. Arguments of the lack of comprehensiveness of admissions tests with respect to true college success notwithstanding, there is a well-documented history on the relationship between admissions tests and performance in college (see Zwick, 2006). Moreover, this research has found admissions test scores to have incremental utility when combined with high school GPA (Sawyer, 2013). Whereas college admission tests have always been used as a component of decisions related to college readiness, the same cannot be said of statewide, standards-based achievement tests. Recent research, however, has begun to examine the extent to which state achievement tests can be used for criterion-related inferences about college readiness.

McGhee (2003) analyzed the predictive capability of the Washington Assessment of Student Learning (WASL) for students enrolled at the five Washington state universities using traditional linear regression methods. When Grade 10 WASL total scores and SAT composite

scores were included in the regression models, both predictors accounted for approximately the same proportion of variance in first-year GPA. However, the WASL did not predict grades in specific college courses as well as the SAT.

Coelen and Berger (2006) analyzed the predictive properties of the Connecticut Academic Performance Test (CAPT) also with respect to the SAT. CAPT Math scores and SAT Quantitative scores were equally potent, accounting for the same proportion of variance in predicting college GPAs when only one or the other was used in the prediction, and neither measure added uniquely to the other when used simultaneously in the prediction. In English language arts, however, students' CAPT scores and SAT Verbal scores accounted for unique proportions of GPA variance when included as simultaneous predictors. While McGhee (2003) found the WASL comparable to the SAT for predicting students' college GPAs, Coelen and Berger concluded the CAPT was limited in this regard as it showed differential prediction depending on institution type and subject area.

In Arizona, Cimetta, D'Agostino, and Levin (2010) investigated whether the state's high school exam could accurately predict students' college GPA at the University of Arizona. Scores from the Mathematics and Writing sections of the exam were significant and positive predictors of first-year GPA. In contrast to what might be expected from the finding by Brown and Conley (2007) of strong content alignment between state standards-based tests in reading and college standards, the Reading section did not contribute to the variance of GPAs explained. Though the results indicated that the high school exam could not replace the SAT, Cimetta et al. (2010) concluded that the state's exam could still serve as a useful indicator of college readiness.

Kingston and Anderson (2013) examined the Kansas State Assessment (KSA) for predicting student success in dual-enrollment college courses, those for which student receive

both high school and college credit. Specifically, high school KSA scores were used to predict the success of high school students taking College Algebra and College English at a community college in Kansas. Results showed that scores from the KSA predicted college course grades as well as the placement tests actually used to place the students into the dual-enrollment courses.

Two studies considered analyses related to the present study's objective of setting benchmarks for readiness. An analysis by D'Agostino and Bonner (2009) focused on the performance levels established by a standard-setting panel on Arizona's high school test and whether they were set at a level suitable to properly define readiness based on GPAs corresponding to B- and C-levels of performance. They found that the standards for proficiency in writing and mathematics were set at or higher than collegiate expectations. However, the proficiency standard in reading appeared to be less rigorous. Other analyses revealed the Reading test provided little useful information about student readiness, consistent with the finding of Cimetta et al. (2010) discussed previously.

Most recently, Nichols-Barrer, Place, Dillon, and Gill (2015) compared the ability of the assessments in the Massachusetts Comprehensive Assessment System (MCAS) and the PARCC consortium to accurately determine whether students will succeed in college. Both the MCAS and PARCC test scores predicted students' first-year GPAs equally well and were comparable to the predictions based on SAT scores. They also found that the CCR benchmark set by the consortium on the PARCC Mathematics assessment predicted a higher level of college performance than the proficiency benchmark set by the state on the MCAS. As proficiency benchmarks were set by states under the No Child Left Behind (NCLB) mandate that every student must be proficient by 2014, many states likely set low standards (see, for example, Bandeira de Mello, Bohrnstedt, Blankenship, & Sherman, 2015), accounting for this difference

between the PARCC and MCAS results. In English language arts, by contrast, the MCAS proficiency benchmark and the PARCC readiness benchmark represented the same level of performance. Nichols-Barrer et al. (2015) indicated that if the MCAS tests were to be used for college readiness interpretations, a higher benchmark would need to be established for the MCAS Mathematics assessment to support its validity for such a purpose.

While D'Agostino and Bonner (2009) and Nichols-Barrer et al. (2015) examined readiness using existing cut scores, other researchers and assessment programs have set benchmarks for readiness using various methods. In Texas for example, benchmarks were set on the state's End-of-Course tests by administering those tests to college students and linking the results to the performance of those same students in the relevant college courses (Texas Education Agency, 2013). Other studies have set college readiness benchmarks indirectly by linking high school assessment scores to scores from a college admissions test, and then identifying the point on the high school test distribution that corresponds to the readiness cut score on the college admissions test (e.g., Furgol, Fina & Welch, 2011; Theaker & Johnson, 2011).

Summary of Research

These studies are part of a growing body of research that examines the use of state tests for college readiness interpretations. Although there are exceptions, the evidence is beginning to show that some assessments may serve as indicators of college readiness on par with traditional admissions tests (e.g., Cimetta et al., 2010; Kingston & Anderson, 2013). This may seem somewhat surprising in light of the presumed difference between factors of student motivation on admissions tests versus standards-based high school tests. More important for the purpose of

the present study is the fact that current research provides an incomplete picture on the validity of using a state's high school assessment for college readiness interpretations.

First, studies that have college outcome data often only provide results from correlation and regression analyses on the predictive capabilities of the high school assessment compared to an admissions test. These studies look at the proportion of variance explained in first-year GPA. From both a test taker's and policymaker's perspective, statistical evidence of this sort may be of little consequence because correlations and regression coefficients have obscure meaning. Unlike college admissions tests where readiness information is inherent in their scales because of their familiarity to these audiences, state achievement tests usually do not share this interpretive meaning. For example, a student could compare an ACT score to a college's ACT average to get a rudimentary sense of whether s/he is a good match for that college. Performance level descriptors that describe what students can do at particular score points provide additional meaning to scales. However, the performance standards on many high school tests often have been defined without any information on the relationship between test scores and performance in college (e.g., D'Agostino & Bonner, 2009).

Just as performance standards are typically set without college outcomes, a shortcoming of several benchmark studies is that readiness benchmarks were set on a state's high school assessment by linking to a college admissions test instead of outcomes measured in college. For example, Furgol et al. (2011) and Theaker and Johnson (2011) likely did so because matched data containing both test scores and college outcomes were unavailable at the time of the investigation. However, interpretations based on assessment scores linked directly to college outcomes provide stronger validity evidence in support of college readiness interpretations. Unlike previous studies, this study sets college readiness benchmarks using students' high school

test scores matched with their actual college outcomes. Although not without their shortcomings (Maruyama, 2012), college readiness benchmarks may prove useful because they quickly convey a student's standing relative to some empirically defined level of performance. They also may be necessary if a state requires a postsecondary readiness level when defining achievement levels on a test and intends to make specific claims about readiness based on test scores.

Method

Data Source

Data for this study were obtained from the Iowa Testing Programs (ITP) and the Office of the Registrar at the University of Iowa and consist of scores on achievement tests taken by students in high school and college transcripts of those students. The high school assessments measure state standards and have been administered annually in the state of Iowa for local use by public and private schools and for use by the state for federal accountability under NCLB. This dataset contained scores from students who graduated from high school in the spring of 2009 and entered the university that same fall. Students' Grade 11 test scores were used because Grade 11 testing is mandated for all students for federal accountability purposes. Data from the Office of the Registrar included transcripts for about 2,000 first-year students who enrolled at the university in the fall of 2009 and who enrolled in specific credit bearing courses that could serve as the operational definition of readiness. The dataset contained course grades, first-year GPA, and ACT scores.

The Iowa Assessments

The Iowa assessments (IAs) are achievement tests aligned to state content standards and designed to identify students' strengths and weaknesses, monitor growth, and predict future performance. The IAs were designed to measure foundation skills and abilities that establish a

basis for continued learning, and predictive relationships have been documented between performance on the IAs and performance in college. For example, Wood and Ansley (2008) found correlations similar to those in previous studies on earlier versions of the IAs (e.g., Loyd, Forsyth, & Hoover, 1980; Qualls & Ansley, 1995). Correlations of scores in specific content areas such as reading or mathematics with first-year GPA ranged from 0.23 to 0.33. The composite score correlation with first-year GPA was equal to 0.39. These results were in line with the correlations observed between sections on the ACT and first-year GPA. In addition, Furgol et al. (2011) used the bivariate distributions of IA and ACT scores in four content areas to determine cut-scores on the IAs that were comparable to the college readiness benchmarks of the ACT. This study adds to this body of research by setting empirical benchmarks on the IAs for Reading, Mathematics, and Science by directly linking test scores to course grades in college. These content areas were selected because they have been and continue to be the focus of federal accountability requirements under NCLB and ESSA.

General Education Courses

The courses selected for setting the benchmarks were part of the general education requirements (GERs) at the University of Iowa. The GERs specify courses for graduation that satisfy distribution requirements in the humanities and in the natural, quantitative, and social sciences, those in the sciences being relevant to this study. If a student took multiple GERs during the first year in one of the natural, quantitative, or social sciences, one course was selected randomly for purposes of the analysis. Random selection from multiple, similar courses is consistent with previous research in setting college readiness benchmarks (e.g. Allen & Sconing, 2005; Allen, 2013).

Course grades in the GERs most closely matched to the content standards of the Mathematics, Science, and Reading assessments were used as the criteria for the two principal objectives of this study. The most common courses for the analysis of the Mathematics test were Math for Business and Math for the Biological Sciences. For the Science test, the most common courses were General Chemistry and Principles of Biology. Courses that satisfied GERs in social science were used for the Reading test, the rationales for this choice being the heavy reading load such courses typically require and the importance of reading for information in the content standards on which the Reading test is based. The most common social science GER was Elementary Psychology. Because not every student took a GER during their first year, the sample sizes used to determine readiness benchmarks differed by subject area and were somewhat smaller than the sample sizes utilizing first-year GPA. There were 1,534 students, 1,421 students, and 1,279 students in the Reading, Mathematics, and Science benchmark datasets, respectively.

Analyses

Objective 1 Bivariate Relationships

In support of Objective 1, the relationship between performance on the IAs and performance in college was examined. It would make little sense to set benchmarks on an assessment if the assessment had a weak relationship with performance in college. Scatterplots of bivariate distributions were examined and correlations between test scores and first-year GPAs were computed. Correlations were also adjusted for differences in variability of the college-enrolled sample relative to the population of Iowa students using the known population variance

of test scores for the Iowa student population and the variance of the sample (Sackett & Yang, 2000).

To better understand the predictive properties of the IAs as well as the extent to which they explain variance in GPA over and above ACT scores and important demographic variables, a hierarchical regression analysis with *a priori* order of entry was conducted. Demographic variables including gender, race (Asian, White, Black, and Hispanic), English language learner (ELL) status, and free- and reduced-lunch (ELL) status were entered first as a combined set of predictors. ACT and IA scores were then added in parallel to examine the unique contribution of each and to determine whether the test added to the model last improved the prediction of first-year GPA. At each step, the change in R^2 was evaluated. Composite scores on the IAs and ACT were used in these analyses because, like first-year GPA, they are broad measures of general achievement and were used in previous criterion-related validation research. A Type I error probability of .01 was used to control for the effects of the multiple hypothesis tests.

Objective 2 Readiness Benchmarks

Objective 2 sought to set college readiness benchmarks on the IAs and evaluate the evidence supporting the benchmarks. To define the benchmarks, grades in typical credit-bearing general education courses were modeled as a function of scores on the IAs using logistic regression (LR). A similar method for determining college readiness benchmarks was used by Allen & Sconing (2005) and Allen (2013), but with a hierarchical component to accommodate the sampling of multiple institutions.

In this study, LR models the probability that a student will receive a given grade in a credit-bearing GER based on the Grade 11 test score. In the present specification, two LR models were defined, one in which students scoring at the cut score have a 50% chance of

earning a B or better in a GER (LR Model B) and the other in which students scoring above the cut have a 50% chance of earning a C or better (LR Model C). Students scoring above the cut score are predicted to have a greater than 50% chance of earning a B (or C) while students scoring below would have a less than 50% chance. The regression model was specified as

$$\log_e \left[\frac{p_i}{1-p_i} \right] = \text{Intercept} + \text{Slope} \times IA_i, \quad (1)$$

where p_i is the probability of success (i.e., a student earns a B or better) given the model and IA_i is the score of student i on the relevant IA test in standard score units. As apparent in Equation 1, the probability that a student is college ready in the defined sense is linearly modeled in terms of log odds, i.e. the log of the ratio of the probability that a student is college ready to the probability that a student is not college ready. Note that we are not arguing that a 50% chance of earning a C is an appropriate definition of success, rather we selected it to demonstrate how such decisions will impact the empirical relationships you are investigating.

After the benchmarks were set, their ability to identify college-ready students was examined in terms of the percents of correct classifications, false positives and false negatives. If the benchmarks do identify a point on the IAs score scale where students more likely to succeed in college, then this fact should be reflected in the actual college outcomes of both groups. The performance of students in college courses who scored above the benchmark was contrasted with those who scored below it in each subject area to examine the ability of the IAs to discriminate successful from unsuccessful students based on the benchmarks.

Finally, the relationship between being college ready in only one or more than one subject area was examined for LR Model B. This analysis used first-year GPA because it is a broader measure of achievement and the appropriate criterion when multiple benchmarks are used together. Whereas each subject area was considered separately when setting the benchmark,

all three subject areas were examined jointly in this analysis to reveal if the benchmarks can be used together to support a validity argument with respect to postsecondary readiness information.

Results

Empirical Evidence of Bivariate Relationships

Descriptive statistics for the sample of students who enrolled at the University of Iowa in the fall of 2009 are provided in Table 1. The performance of this sample on the IAs is higher than that of the full Iowa student population. In that population, the average IA standard score in Grade 11 is about 25-30 points lower than the means observed for these scores in Table 1. Note that to be included in the analytic sample, students had been admitted to the University of Iowa, opted to take the ACT, and were required to have scores on the IAs in Reading, Mathematics, and Science. Also note that because the composite score on the IAs requires additional assessments, there were fewer students with composite scores for this variable.

The relationship between course grades and IA test scores is illustrated in the graphs provided in Figure 1. It is apparent that higher course grades are associated with higher test scores. In Reading and Science, the progressively higher averages moving up the grade scale suggest a somewhat stronger relationship between college grades and test scores in these areas than in Mathematics. The more gradual increases in Mathematics might be due to the fact that students may take higher level math courses to satisfy the quantitative GERs.

The correlation between the IAs and a traditional admissions test provides additional empirical evidence in support of the IAs as indicators of college readiness. In the population of ACT test takers in the state of Iowa, the correlation between scores on the IAs and ACT scores for Reading, Mathematics, and Science were 0.75, 0.76, and 0.68, respectively (see Furgol et al.,

2011). As these correlations have not been adjusted for restriction of range (not every Iowa student takes the ACT), they represent lower bounds of the true relationship between the two measures. In addition, the correlations between test scores and first-year GPAs for the IAs and the ACT are presented in Table 2. The unadjusted correlations between test scores and first-year GPAs were higher for the ACT and typical of the magnitudes of correlations between predictors and college success criteria in selected samples. When adjusted for range restriction, however, the correlations for the IAs were higher than those for the ACT.

To further examine the predictive capabilities of the IAs, the results of a hierarchical regression analysis with *a priori* order of entry are presented in Table 3. Only the R^2 , ΔR^2 , and associated p-value are presented here because of interest in this analysis is establishing the relevance of tests scores over and above that of background characteristics of students. After controlling for demographic variables including gender, race, ELL status and FRL status, both the IAs and the ACT Composite score yielded a statistically significant increment in predicted first-year GPA (see Step 2 of Table 3). In Step 3, the composite scores for both the IAs and the ACT yielded a statistically significant increment in R^2 over the other test, although those increments only explained about an additional 1% of the variance in first-year GPA. This implies that, at this university, the IAs and the ACT have an equal ability to predict first-year GPA and little unique explanatory information.

Readiness Benchmarks

Table 4 provides the cumulative frequency distribution of course grades in the samples used to set the college readiness benchmarks. As observed by Allen and Sconing (2005) in a study of many higher education institutions, about 50% of the students in this sample earned

letter grades of B or better in social science and quantitative GERs. The percentage in natural science GERs was only slightly lower at 44. Note that only about 12 percent of the students earned letter grades below a C- and by far the majority of grades received by students in the selected GER courses were passing grades.

The results from the logistic regression analyses are summarized in Table 6, which reports for each LR model and subject area assessment the parameter estimates, their standard errors and the standardized difference for testing whether the parameter estimate equals zero. The parameter estimates are expressed in the metric of the standard score scale of the IAs, so the slope coefficients appear small in magnitude. All estimated intercepts and slopes were significantly different from zero ($p < .01$), an expected result given the large sample sizes.

The readiness benchmark cut-scores derived from the logistic regression results reduce to the absolute value of the ratio of the intercept to the slope in Table 5. The resulting college readiness benchmarks for Reading, Mathematics, and Science were 317, 319, and 336, respectively, for LR Model B and 243, 248, and 256 for LR Model C. These cut scores are the IA standard scores that, when substituted into the LR model, are associated with a .50 probability of earning a letter grade of B (or C) or better in the GER courses at the University of Iowa. Figure 2 plots the model-based probabilities of earning the criterion grade or better in each subject area across the full range of IA standard scores. Similarly shaped curves were observed in the three subject areas. Where each curve crosses the .50 threshold corresponds to the benchmark for that subject area and criterion grade. The probabilities represented in these figures may represent important information for students and other stakeholders compared to the dichotomous indicator offered by cut scores alone (Maruyama, 2012).

The fitted models in Figure 2 show how higher test scores correspond to greater likelihood of success in postsecondary education. At the low end of the IA score scale, students were more than twice as likely to get a C or above using LR Model C than they were to get a B or above using LR Model B. At the upper end of the IA score distribution, students were very likely to pass a course (over .95 probability of earning at least a C letter grade). However, the probability of those same students earning a B or above was approximately .80. As there are many factors influencing a student's grade, it is not surprising that the probability estimated by either LR model never equals 1.0.

The classification consistency of the LR models can be described by the relative frequencies of correct and incorrect model-based decisions. Table 6 provides the percentage of correct classifications, false-positive errors, and false-negative errors for the two fitted LR models. A false-positive error indicates a student was deemed college-ready based on their IA score but earned a letter grade below a B (or C) in the corresponding GER course. Similarly, a false-negative error indicates that a student was deemed not college-ready based on their IA score but earned a letter grade equal to a B (or C) or above in the corresponding GER. The percentage of correct classifications for LR Model B are nearly identical to those found by Allen and Scoring (2005) and are about 63 percent. False positives tend to be more frequent than false negatives, and the classification consistency results suggest that the benchmarks have greater utility for IA scores in the top and bottom quarter of the score distributions.

The extent to which the benchmarks separated students into a college-ready group and a not college-ready group was examined next. Table 7 summarizes the performance of students in GER courses who met the benchmarks of the two LR models and those who did not. In Reading and Mathematics, 61 percent of the students who were classified as college-ready by LR Model

B earned a grade of B or better, and about 88 percent of the students who were classified as college-ready by LR Model C earned a C or better. Of the students who scored above the Science benchmark of LR Model B, 57 percent earned a grade of B or better; the same figure for LR Model C was 90 percent. Among students who did not meet the benchmark in any subject area, over 70 percent still attained a letter grade of C or better in the GER courses analyzed. However, those same students (i.e., those not meeting any benchmark) were much less likely to earn a B or above. These results are not surprising given the fact reported previously that relatively small percentages of student received letter grades below C in the GER courses.

The results in Table 7 suggest that the college readiness benchmarks set on the IAs are potentially useful indicators of readiness to achieve a certain grade in a credit-bearing course at this institution. Note that scoring above a benchmark set by a given model does not guarantee a given course grade in the GERs. Only about 60 percent of the students who scored above the benchmarks of LR Model B earned a B or above. However, the usefulness of the model-based benchmarks is supported by the fact that students who met the benchmark were almost twice as likely to earn a B as students who did not meet the benchmark for a given subject area. Another important observation from the results in Table 7 concerns the relative discrimination provided by LR Models B and C between those meeting and not meeting the CCR benchmark. Greater discrimination exists when the LR criterion grade matches the course grade of interest. That is, if the operational definition of CCR is a letter grade of B or better, then LR Model B provides the better discrimination, whereas if the operational definition is C or better, then LR Model C provides better discrimination. In other words, the definition of the CCR criterion matters for model specification and use in practice.

Finally, we examined the relationship between the number of CCR benchmarks met and overall performance in the first year of college. We expected that students who met the benchmark in all three subject areas would have the highest average first-year GPA, and this average would decrease as the number of subject areas in which a student was ready decreased. Students who did not meet the benchmark in any subject area were expected to have the lowest average first-year GPA. Results are provided in the bar graphs of Figure 3 and are consistent with expectations. Students who did not attain any benchmark (21 percent of the sample) had the lowest average first-year GPA of 2.74. For students who met all three benchmarks (35 percent of the sample), the average first-year GPA was 3.28, approximately equal to a B+. There were 21 percent of the students who met one benchmark and 23 percent who met two benchmarks. Students who met two benchmarks had an average first-year GPA of 3.04, slightly higher than a B average. Among students who attained only one benchmark, the average first-year GPA across all three subject areas was 2.91.

Discussion

Receiving a high school diploma does not guarantee that students have the necessary knowledge and skills required to be successful in postsecondary education. To be successful in college requires a variety of skills and academic behaviors, in addition to content knowledge learned in the classroom (see Conley, 2007). However, the academic knowledge and skills of students entering college are fundamental to their eventual success. Given that 29 to 41 percent of the students at public post-secondary institutions report taking at least one remedial course in college (NCES, 2016), providing college readiness information to high school students is one way to address high remediation rates and encourages students to seek additional instruction if needed.

This study attempts to demonstrate how a state's standards-based achievement test can support college readiness interpretations for high school students. A variety of empirical evidence was presented that summarized the relationship between performance on the assessment and performance in college. Higher scores on the high school assessment were associated with higher grades in college courses and a higher first-year GPA. Previous studies have found similar relationships between the IAs and performance in college (e.g., Wood & Ansley, 2008), which suggest that the relationship between performance on the IAs and college outcomes is stable.

The benchmarks defined here offer a simple way to indicate whether students are ready for postsecondary education based on definitions of letter grades in credit-bearing course. The strong positive relationships observed between the assessment scores and performance in college supported the development of CCR benchmarks for the IAs. They represent the level of achievement a student needs to succeed in GER courses. The benchmarks are useful for identifying deficiencies in specific academic areas given that students who met a benchmark were more likely to earn letter grades of B (or C) or above in the corresponding GER courses than those who scored below the benchmark. In addition, as the number of CCR benchmarks students attained increased, so did the first-year GPA.

The correlations between the state assessment and the traditional admissions test, as well as the similar pattern of correlations seen among the assessments and first-year GPA, show that the knowledge, skills, and abilities measured by both tests are related. No claim is made that the knowledge, skills, and abilities are the same, only that they are similar. Brown and Conley (2007) analyzed the content of 20 state high school exams relative to a set of CCR standards, finding that the tests were moderately aligned with a subset of the standards, although the

relationship varied by content area. They also found that the range of knowledge assessed by the items on the state tests did not cover the full range of objectives in the CCR standards. Given the large variety of high school assessments used across nation, it is not surprising that Brown and Conley (2007) found mixed results. This suggests that the content alignment of tests is but a small piece in a comprehensive approach to understanding what it means to be ready for postsecondary education. Camara (2013) further argues that alignment studies and other judgmental approaches are insufficient for supporting college readiness interpretations, which require empirical evidence like that presented here.

Through the setting of benchmarks, we demonstrated how a state assessment can serve as a CCR indicator and provide readiness information to high school students. There are broader implications as well. First, a state's standards-based assessment can monitor the college readiness of students statewide because such a test is typically administered to all students in at least one high school grade, whereas results from a college admissions test might represent a self-selected sample.

Second, reporting of CCR information to the entire student population can inspire some students to attend college who otherwise might not have had college aspirations. School counselors or practitioners might also use this information. For example, a student who met the CCR benchmarks in Reading and Science but not Mathematics might be advised to take an additional math course given their current readiness standing and the fact that the benchmarks had low false negative percentages.

Third, further utility could come from postsecondary institutions examining if state assessment scores provide information for placement or admission purposes, as other researchers have considered with promising results (e.g., Kingston & Anderson, 2013). Although the

accuracy of the CCR benchmarks was acceptable for use as one indicator of college readiness, new uses would require further validation research because the consequences of test score use will have a larger role in a validity argument as the stakes associated with a proposed interpretation increase (Kane, 2013).

It should be noted that logistic regression has been commonly used to set college readiness benchmarks (e.g., Allen, 2013; Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011). Standard setting methods are also of interest in this regard (McClarty, Way, Porter, Beimers, & Miles, 2013). The benchmarks proposed here were developed for reporting CCR information to students who might not otherwise receive such information. They are easily interpretable as they represent the test score that is associated with at least a 50% chance of earning a letter grade of interest in GER courses. If they were to be used for placement decisions or in admissions context, a more thorough study that evaluated multiple cut scores would be required to more accurately determine the “best” cut score. This process could consider results from a Receiver Operating Characteristic (ROC) curve analysis that examines the sensitivity, specificity, false-negative errors, and false-positive errors associated with different cut scores. This process would also carefully weigh the consequences of false-positive and false-negative errors, as the appropriate balance will vary by context.

Most CCR benchmarks for college admissions tests define readiness in terms of a typical student at a typical college (e.g., Allen, 2013). The benchmarks are not institution specific. In this study, readiness was defined for a specific institution. While this provides more accurate and relevant information to applicants, future research should be directed at determining the extent to which the results in this paper generalize. Institution type (e.g. community college, selective 4-year liberal arts college, mid-size public university) may be relevant to the generalizability of the

CCR benchmarks (Porter & Polikoff, 2012). The methods of validity generalization should also be considered (Linn, Harnisch, & Dunbar, 1981).

A second limit on the potential utility of CCR benchmarks involves the perennially elusive criterion of success in education and how it is defined and operationalized. This study defined readiness based on grades in specific GER courses and test scores because of their ready availability and property as unobtrusive measures given that all students at some point take a state test in high school and will also have a university record. However it is possible that other variables might have a mediating or moderating effect on readiness. Studies of this sort, interested ultimately in the validation of a CCR claim or inference, should include other measures in conjunction with test scores.

Several recent studies have laid the groundwork for a multifaceted approach to measuring college readiness for which Conley (2007) and other researchers have been advocating. Robbins, Allen, Casillas, Peterson, and Le (2006) identified several non-cognitive attributes that are associated with success in college. Most recently, Gaertner and McClarty (2015) demonstrated how non-cognitive measures could be included in a college readiness index and how these measures could be used to provide CCR information to middle school students. In their work, first-year GPA was predicted equally well using Grade 11 SAT or ACT scores and using achievement measures, behavior and other non-cognitive measures available in middle school, although not as well as SAT/ACT score and high school GPA combined. Mattern, Allen, and Camara (2016) highlighted this issue when they discussed how the congruence between indicators of college readiness and criteria of college success will impact conclusions made from a study.

The term readiness is used in this article for two reasons. First, readiness and preparedness are used interchangeably in most social science disciplines (Camara, 2013). Second, if the factors and behaviors in the non-cognitive realm called for by some researchers were to be included in the definition of college readiness (e.g., Conley, 2007; Gaertner & McClarty, 2015), a similar empirical study would need to be undertaken to establish the relationship between the alternative measure(s) and performance in college. Regardless of how college readiness is operationalized and defined, this study has demonstrated the types of empirical relationships that must exist between measures if CCR interpretations are to be made from test scores. While the primary focus of our study was on criterion-related validity evidence, other sources of validity evidence are required for a comprehensive validity argument. Another key source is content-related validity evidence, as described in a general validation framework for college readiness interpretations outlined by several researchers (Camara, 2013; Welch & Dunbar, 2011). Content-related validity evidence comes from the match between the standards, the assessment, and the resulting interpretations. Because sound test development processes are intricately tied to content-related validity evidence (Mislevy & Haertel, 2006; Welch & Dunbar, 2011), evidence must be collected throughout the process in support of the validity argument.

This study contributes to the growing body of research examining the relationship between state assessments and postsecondary success. It illustrated the type of empirical study that any testing program making similar readiness interpretations must undertake. By demonstrating how the IAs can be a potentially useful indicator of college readiness, it makes the case that scores from any standards-based achievement test can be validated for CCR interpretations. These interpretations could be extended down to earlier grades by linking through a vertical scale (Furgol et al., 2011) or using a different type of growth model (Fina,

2014). Such solutions might be cost-effective alternatives and widely applicable because additional information is gleaned from an existing assessment program. States should recognize the opportunities afforded by their own assessment systems as they consider the costs and benefits of the next-generation CCR assessments.

References

- ACT. (2014). *ACT technical manual*. Iowa city, IA: ACT, Inc.
- Allen, J., & Sconing, J. (2005). *Using ACT assessment scores to set benchmarks for college readiness* (No. 2005-3). Iowa City, IA: ACT, Inc.
- Allen, J. (2013). *Updating the ACT college readiness benchmarks* (No. 2013-6). Iowa City, IA: ACT, Inc.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). *Mapping state proficiency standards onto NAEP scales: Results from the 2013 NAEP reading and mathematics assessments* (NCES 2015-046). U.S. Department of Education, Washington, DC: National Center for Education Statistics.
- Brown, R. S., & Conley, D. T. (2007). Comparing state high school assessments to standards for success in entry-level university courses. *Educational Assessment, 12*, 137–160.
- Camara, W. J. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice 32*(4), 16-27.
- Cimetta, A. D., D'Agostino, J. V., & Levin, J. R. (2010). Can high school achievement tests serve to select college students? *Educational Measurement: Issues and Practice, 29*(2), 3-12.
- Coelen, S. P., & Berger, J. B. (2006). *First steps: An evaluation of the success of Connecticut students beyond high school*. Quincy, MA: Nellie Mae Education Foundation.

- Conley, D. T. (2007). *Redefining college readiness*. Eugene, OR: Educational Policy Improvement Center.
- D'Agostino, J. V., & Bonner, S. M. (2009). High school exit exam scores and university performance. *Educational Assessment, 14*, 25-37.
- Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015).
- Fina, A. D. (2014). *Growth and the college readiness of Iowa students: A longitudinal study linking growth to college outcomes*. (Doctoral thesis). Iowa City, IA: University of Iowa.
- Furgol, K., Fina, A. D., & Welch, C. (2011). *Establishing validity evidence to assess college readiness through a vertical scale*. Paper presented at the AERA Annual Meeting, New Orleans, LA.
- Gaertner, M., & McClarty, K. L. (2015). Performance, perseverance, and the full picture of college readiness. *Educational Measurement: Issues and Practice, 34*(2), 20–33.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*: 1-73
- Kingston, N. M. & Anderson, G. (2013). Using state assessments for predicting student success in dual-enrollment college classes. *Educational Measurement: Issues and Practice, 32*: 3–10.
- Linn, R. L., Harnisch, D. L., Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first-year grades in law school. *Applied Psychological Measurement, 5*(3), 281-289.
- Loyd, B. H., Forsyth, R. A., & Hoover, H. D. (1980). Relationship of elementary and secondary school achievement test scores to later academic success. *Educational and Psychological Measurement, 40*(4), 1117-1124.

- Maruyama, G. (2012). Assessing college readiness: Should we be satisfied with ACT or other threshold scores? *Educational Researcher*, 41(7), 252-261.
- Mattern, K., Allen, J., & Camara, W. (2016). Thoughts on a multidimensional middle school index of college readiness. *Educational Measurement: Issues and Practice*, 35(3), 30–34.
- McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, 42(2), 78-88.
- McGhee, D. E. (2003). *The relationship between WASL scores and performance in the first year of university*. Seattle, WA: Office of Educational Assessment, University of Washington.
- Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, 25, 6–20.
- National Center for Education Statistics. (2016). *Web tables: Profile of undergraduate students: 2011-12*. (NCES 2015-167). Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.
- Nichols-Barrer, I., Place, K., Dillion, E., & Gill, B. (2015). *Predictive validity of MCAS and PARCC: Comparing 10th grade MCAS tests to PARCC Integrated Math II, Algebra II, and 10th grade English Language Arts tests*. Boston, MA: Mathematica Policy Research.
- Porter, A. C., & Polikoff, M. S. (2012). Measuring academic readiness for college. *Educational Policy*, 26(3), 394-417.
- Qualls, A. L., & Ansley, T. N. (1995). The predictive relationship of ITBS and ITED to measures of academic success. *Educational and Psychological Measurement*, 55(3), 485-498.

- Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*(3), 598–616.
- Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admission decisions. *Applied Measurement in Education, 26*(2), 89- 112.
- Sackett, P.R., & Yang, H. (2000). Correction for range restriction: an expanded typology. *Journal of Applied Psychology, 85*(1), 112-118.
- Texas Education Agency. (2013). State of Texas Assessments of Academic Readiness standard setting technical report. Retrieved March, 2017 from <http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769804117&libID=25769804117>
- Theaker, R. & Johnson, C. (2011). College readiness linking study: A study of the alignment of the RIT Scales on NWEA’s MAP assessments with the College Readiness Benchmarks of EXPLORE, PLAN and ACT. Portland, OR: Northwest Evaluation Association.
- Viadero, D. (2017). Assessments: A snapshot of a field in motion. *Education Week, 36*(2), 5.
- Welch, C., & Dunbar, S. B. (2011). *K–12 assessments and college readiness: Necessary validity evidence for educators, teachers and parents*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Wyatt, J., Kobrin, J., Wiley, A., Camara, W., & Proestler, N. (2011). *SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance*. New York, NY: College Board.

Wood, S., & Ansley, T. (2008). *An investigation of the validity of standardized achievement tests for predicting first-year college GPA and college entrance exam scores*. Paper presented at the annual meeting of the NCME, New York.

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

Table 1

Descriptive Statistics for the Sample

Measure	Variable	N	Mean	Standard Deviation
IA	Reading	1814	323.97	32.0
	Math	1814	322.77	23.7
	Science	1814	333.50	28.2
	Composite	1493	317.54	21.4
ACT	Reading	1814	25.71	5.0
	Math	1814	24.58	4.3
	Science	1814	24.50	3.9
	Composite	1493	25.14	3.8
GPA	First-Year	1814	3.04	0.7

Table 2

Correlations with First-Year Grade Point Averages

Subject	Iowa		ACT	
	Unadjusted	Adjusted	Unadjusted	Adjusted
Reading	0.284	0.351	0.309	0.342
Mathematics	0.297	0.441	0.364	0.399
Science	0.322	0.446	0.307	0.341
Composite	0.371	0.513	0.394	0.450

Table 3

Results of Hierarchical Regression Analyses

Variable Set Entered	R^2	ΔR^2	$\Delta R^2 p$
Step 1			
Demographics	0.020	0.020	<.001
Step 2			
i. Demographics	0.158	0.138	<.001
Iowa Core Total			
ii. Demographics	0.167	0.148	<.001
ACT Composite			
Step 3			
i. Demographics	0.172	0.014	<.001
Iowa Core Total			
ACT Composite			
ii. Demographics	0.172	0.005	.002
ACT Composite			
Iowa Core Total			

Note: Variables entered at each step are in bold.

Table 4

Cumulative Frequency Distributions of Letter Grades

in General Education Courses

	Reading	Math	Science
A+	100.0	100.0	100.0
A	97.9	96.5	98.2
A-	87.3	82.5	87.5
B+	78.8	75.9	82.9
B	68.8	66.4	74.8
B-	51.3	48.8	55.7
C+	41.0	41.2	48.6
C	32.5	31.4	37.7
C-	18.6	17.9	17.6
D+	11.8	13.4	10.5
D	8.6	10.8	7.9
D-	5.0	6.3	3.6
F	3.3	5.1	2.0

Table 5

Estimates from the Logistic Regression Models

Subject Area	LR Model	Parameter	Estimate	Standard Error	Est./S.E.
Reading	B	Slope	0.021	0.002	10.50*
		Intercept	-6.661	0.581	11.46*
Mathematics	B	Slope	0.025	0.003	8.33*
		Intercept	-7.977	0.858	9.30*
Science	B	Slope	0.028	0.002	14.00*
		Intercept	-9.414	0.820	11.48*
Reading	C	Slope	0.021	0.002	9.25*
		Intercept	-5.115	0.703	7.27*
Mathematics	C	Slope	0.021	0.003	7.10*
		Intercept	-5.196	0.940	5.53*
Science	C	Slope	0.024	0.003	8.92*
		Intercept	-6.131	0.851	7.20*

* $p < .01$.

Table 6
 Classification Consistency Results for the
 College Readiness Benchmarks

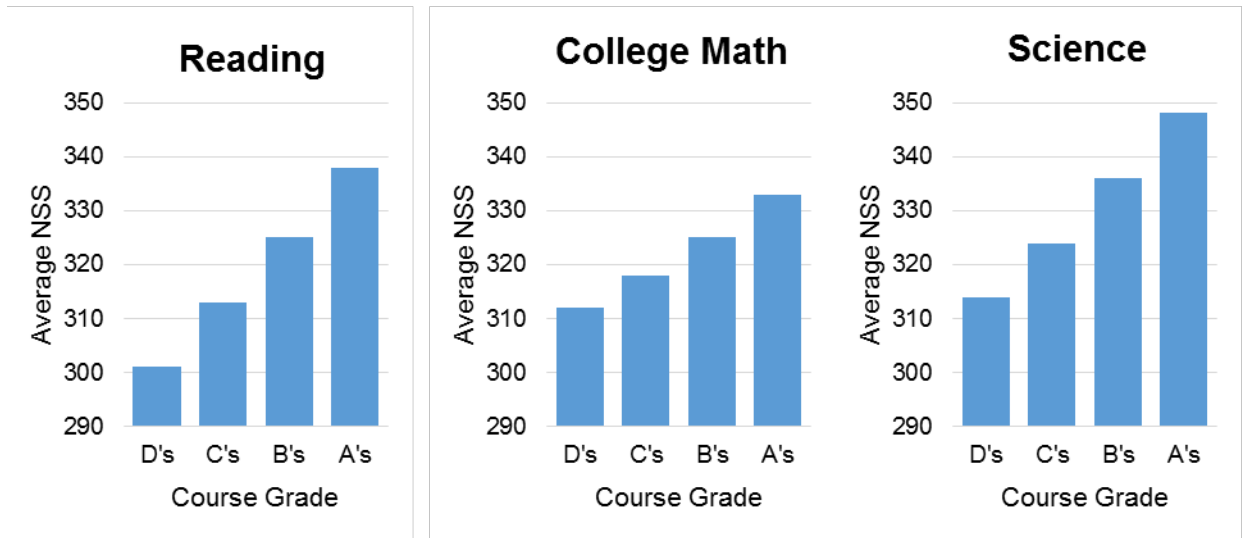
Subject Area	Model Criterion	Decision Percents		PCC
		False Positive	False Negative	
Reading	B	20.8	16.3	62.9
Mathematics	B	23.6	14.9	61.6
Science	B	22.4	14.2	63.4
Reading	C	18.5	0.5	81.0
Mathematics	C	17.7	0.5	81.8
Science	C	16.7	0.8	82.5

Note: PCC = percentage of correct classifications.

Table 7

Percentage of Students at or above Certain Letter Grades in Each Subject Area

Subject	Logistic Criterion	Benchmark	Did not Meet Benchmark		Met Benchmark	
			C or Above	B or Above	C or Above	B or Above
Reading		317	73	35	89	61
Math	50% B	319	75	37	87	91
Science		336	74	30	90	57
Reading		243	73	45	81	49
Math	50% C	248	64	18	82	51
Science		256	48	10	83	45



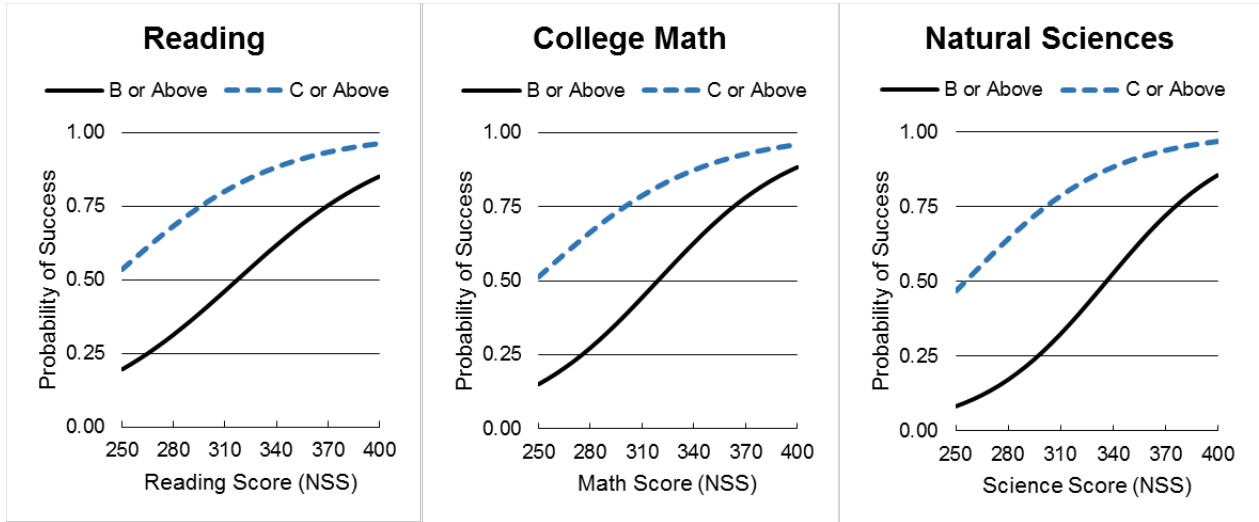
(a)

(b)

(c)

Figure 1.

Average Standard Score on the IAs by course grade.



(a)

(b)

(c)

Figure 2.

Likelihood of success in each subject area.

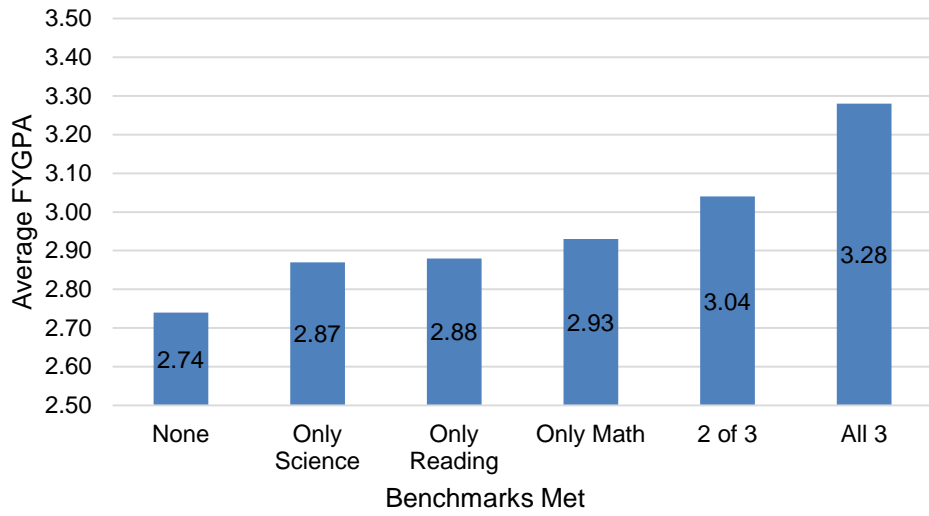


Figure 3.

Average first-year GPA by number of benchmarks met for LR Model B.