Chapter 2
Methodology and Limitations

This chapter presents the methodology and limitations of the study. We first explain why each of the case study sites was selected. We then describe the college admissions exams common to each case study site. This is followed by a discussion of how we created the coding categories and standards used to evaluate test alignment. The chapter concludes with the limitations of our study.

Selection of the Case Study Sites

The selection of the case study sites is best understood within the larger context of the Bridge Project. The Bridge Project is currently exploring whether the lack of compatibility between high school and higher education policies and practices hinders student transitions from secondary to postsecondary education. Of particular importance is whether traditionally underrepresented and economically disadvantaged students who have fewer resources (both information as well as financial) at their disposal also have fewer opportunities to learn about college.

The case study sites were chosen for a variety of state-specific reasons; however, most of these 5 states are actively involved in changing their policies that relate to student transitions between high school and college. The state-specific reasons are as follows:

Texas

In 1996, the 5th Circuit Court of Appeals ruled in its Hopwood decree that all public universities must stop using race/ethnicity as a variable in their admission decisions. In reaction to the ruling, in the 1997 session, the Legislature passed the Top Ten Percent Rule, stating that all Texas seniors in the top ten percent of their high school classes can choose which public university they would like to attend; the universities had to admit any student in the top ten percent of her/his class. Project researchers were interested in whether the court decision changed lower-income and traditionally underrepresented students' motivation to learn about higher education opportunities.

Another area of interest is the lack of clear differences within and between the two main university systems, the University of Texas and Texas A&M. Despite being from the same system, University of Texas institutions do not necessarily share common characteristics or policies. Likewise, each branch of the Texas A&M system has its own guidelines. Additionally, differences between the University of Texas and Texas A&M systems themselves are not readily apparent. Project researchers wanted to explore whether students are confused by the lack of structure within and between these two systems, and what signals the institutions send students.

Finally, Texas has a history of a large achievement gap between white, non-Latino students, African-American students, and Latino students. Project researchers wanted to learn if there were similar patterns of gaps in student knowledge about opportunities to learn about college.

California

Like Texas, California's public institutions were struggling with the aftermath of policy changes (SP1 and 2 and Proposition 209) that halted the use of affirmative action in public higher education admissions decisions. And, like Texas, California is a diverse state with large achievement gap problems. In stark contrast to Texas, however, California has highly structured and tiered postsecondary systems, as evidenced by its delineated UC, CSU, and community college segments. The UC and CSU systems, for example, can be clearly differentiated with respect to structure, purpose, and admissions requirements (as set forth in the form of a-g course requirements). Researchers wanted to know if California students had more knowledge of institutional and system differences, and of course requirements than students in the other case study sites, particularly Texas.

Georgia

Georgia is known as a pioneer in the area of P-16 reform. P-16 reform consists of a coalition among members from postsecondary education, elementary and secondary education, youth advocacy groups, the private sector, technical institutes, and the local community that aims to ease student transitions from secondary to postsecondary

education through a variety of diverse means including outreach programs to disadvantaged students and professional development programs for teachers. In addition, Georgia has the HOPE scholarship, geared toward keeping high-achieving Georgia students in state for college. Researchers wanted to learn if and how the councils have changed the policy signaling process for secondary students, and if the HOPE Scholarship has influenced student aspirations and motivation to learn about college opportunities.

Maryland

Like Georgia, Maryland was one of the first states that developed a state-level K-16 council. Analogous to Georgia's P-16 councils, K-16 councils in Maryland oversee student transition from secondary to postsecondary instruction and teacher preparation-related policies. Researchers wanted to learn if policy development and signaling processes had changed, given the involvement of the council. Additionally, Maryland also had large achievement gap issues, particularly between white, non-Latino students and African American students, and researchers wanted to learn if those differences appeared with respect to student knowledge about college opportunities.

Oregon

Oregon is one of pioneers in the standards and assessment movement and has sustained its efforts for over a decade. In 1991, the Oregon legislature mandated certificates of mastery as part of its overall education reform plan of revised standards and assessments. In 1993, in reaction to the Certificate of Initial Mastery (CIM) and Certificate of Advanced Mastery (CAM) for high school students, the Oregon University System (OUS) began developing the Proficiency-Based Admission Standards System (PASS). Fearing that the CIM and CAM would lower the level of high schools' curricula, the OUS defined the knowledge and skills necessary for students to enter into, and succeed at, the OUS institutions. PASS is proficiency-based, and utilizes teacher judgment in the rating process. Researchers wanted to understand the mandated Oregon reforms, the voluntary PASS system, how they are being implemented in the schools,

how they relate to one another, and how disadvantaged secondary students, make sense of the constantly evolving high school exit and college entrance policy environments.

Assessments Examined

Because of the magnitude of tests available, it was necessary to limit the number of tests examined in this study. We restricted our analysis to math and English/Language Arts (ELA) measures administered to high-school and incoming first-year college students. No assessments at the postsecondary level were examined. We chose math and ELA because most remediation decisions at the postsecondary level are based on achievement deficiencies in these areas.

The assessments were limited to those used by selected institutions in our five case study sites. Because the kinds of tests administered may vary by college, it is important to sample exams from a range of institutions. Namely, the minimum skill level required of students entering a highly selective institution may differ from the level required of students entering a less selective college. As a result, the content of remedial college placement tests used to assess entry-level skills can vary by institution. For each of our sites, we examined assessments administered by colleges that represented a range from less selective to highly selective. However, the chosen institutions are not a scientific sample.

Below we provide more details of the college admissions assessments that are used nationally, and are included in all of our case study sites. State-specific tests are described in each of our case study chapters.

National College Admissions Tests Used in Each Case Study Site

The first set of tests we examined, which includes the SAT I, SAT II, ACT, and AP exams are used in our five case study sites, as well nationally, to aid in college admissions decisions. For those students applying to a four-year institution, many are required to take either the SAT I or ACT, and, at more selective schools, several SAT II exams as part of the admissions process. While the AP tests are not a requirement, admissions officers are likely to view students with AP experience as better-prepared and more competitive applicants.

The SAT I, a three-hour mostly multiple-choice exam, is intended to help admissions officers distinguish applicants more qualified for college-level work from those less qualified. It is not designed to measure knowledge from any specific high school course, but instead measures general mathematical and verbal reasoning. The SAT II is a series of one-hour, mostly multiple-choice tests that assess in-depth knowledge of a particular subject, and is used by admissions officers as an additional measure with which to evaluate student subject-matter competence. The SAT II is used primarily at the more selective institutions and is taken by far fewer students than is the SAT I. For this study, we examined the following SAT II tests: Mathematics Level IC, Mathematics Level IIC, Literature, and Writing. The SAT II Mathematics Level IC test assesses math knowledge commonly taught in three years of college preparatory math courses, whereas the SAT II Mathematics Level IIC test assesses math knowledge in more than three years of college preparatory math courses. The SAT II Literature test assesses students' proficiency in understanding and interpreting reading passages, and the SAT II Writing test assesses students' knowledge of standard written English.

The ACT is an approximately three-hour exam consisting entirely of multiple-choice items. Developed to be an alternative measure to the SAT I in evaluating applicants' chances of success in college, it does not emphasize general reasoning (as does the SAT I) but is instead a curriculum-based exam that assesses achievement in science, reading, language arts, and math (Wightman & Jaeger, 1988). We include only the reading, language arts, and math sections for this study. The AP tests are used to measure college-level achievement in several subjects, and to award academic credit to students who demonstrate college-level proficiency. We examine the AP Language and Composition exam for this study.[1]

Coding Categories

Two raters examined alignment among the different types of assessments using several coding categories, which are described below.

[1] We exclude the two AP exams in calculus (i.e., Calculus AB and Calculus BC) because they are markedly different from the other studied math tests. For example, they do not include material from any other mathematical content area except calculus, and are the only measures that require a graphing calculator.

The coding categories for both math and ELA describe the technical features, cognitive demands, and content of each assessment. The technical features category involves characteristics such as time limit and format. The cognitive demands category captures the kinds of cognitive processes elicited. For reasons that will be explained in another section, the content category is slightly different for math and ELA. In math, the content category captures what is being assessed (i.e., particular content area such as elementary algebra or geometry). In ELA, the content category describes the reading passage.

The cognitive demands category and the content category in math are the focus of this study because discrepancies in these areas can potentially send mixed messages to students about the kinds of skills they should learn in order to be prepared for college-level courses. Although variations in technical features and in the ELA content category are believed to have less direct impact on the kinds of signals students receive, it is nevertheless important to document discrepancies in these areas. Technical features, such as test format, can facilitate or limit the kinds of skills that are measured (Bennet & Ward, 1993). We describe differences in the ELA content category to fully characterize the range of test content possible. Coding categories for each subject appear in Appendix A, and will be described in more details shortly.

We created the above coding categories by exploring different ways of summarizing test content. We examined several sources, including test frameworks, such as those used to develop the National Assessment of Educational Progress (NAEP), as well as coding categories used in previous studies of alignment (Education Trust, 1999; Kenney, Silver, Alacaci, & Zawojewski, 1998; Webb, 1999). We then combined and modified the different sources to produce coding categories that addressed the range of topics and formats appearing on the tests included in this study.

Math Coding Categories

The first aspect of the math coding categories concerns technical features. Technical features describe test length, time limit, format, and characteristics that can be described by inspection of test instructions or items. In math, items could be classified as one of four formats: multiple-choice, quantitative comparison, grid-in, or open-ended.

Multiple-choice items require students to select their answer from a list of possible options. Quantitative comparison items require students to determine the relative sizes of two quantities. Because quantitative comparison items ask students to select their answer from four possible options, they are considered a subset of the multiple-choice format. However, we distinguish between the two types of formats because response options across multiple-choice items vary from one question to the next, whereas response options across quantitative comparison items remain the same. Grid-in items require students to produce their own answer and mark their answer in a corresponding grid. Open-ended items also require students to produce their own answer, but differ from grid-in items in that the former item type can take on negative values. Additionally, many of the open-ended items in our study require students to explain their reasoning.

Technical features also include characteristics that can be described by examining the test instructions or items. These include characteristics such as provisions for the use of tools such as calculators or rulers, the use of diagrams or other graphics, the use of formulas, whether formulas were provided or had to be memorized, and whether each item was contextualized (i.e., a word problem that made reference to a real-life situation). The use of formulas was sometimes difficult to determine because problems can be solved in multiple ways, and in some cases an item could be solved either with or without a formula. Items were coded as requiring a formula only if it was determined that the formula was necessary for solving the problem.

For the content category, we listed several math content areas, ranging from basic through advanced math. The content areas included prealgebra, elementary algebra, intermediate algebra, planar geometry, coordinate geometry, statistics and probability, trigonometry, and miscellaneous. Almost all of the tests we examined had specifications that included many or all of these content areas. We listed subareas as means of making the distinctions among the main content areas clearer, but raters coded using only the main content areas (see Appendix A for the list of subareas).

To evaluate the cognitive demands of a test, we needed a coding scheme that captured different levels of cognitive processes, from routine procedures to complex problem solving. This led us to adopt the same criteria as those used for NAEP, namely

conceptual understanding, procedural knowledge, and problem solving. The descriptions of each are described in Table 2.1 below.

Table 2.1
Descriptions of the Math Cognitive Demands Category

Cognitive Process	Definition
Conceptual understanding	Reflects a student's ability to reason in settings involving the careful application of concept definitions, relations, or representations of either
Procedural knowledge	Includes the various numerical algorithms in mathematics that have been created as tools to meet specific needs efficiently
Problem solving	Requires students to connect all of their mathematical knowledge of concepts, procedures, reasoning, and communication/representational skills in confronting new situations

Source: Mathematics Framework for the 1996 and 2000 National Assessment of Educational Progress (NAGB, 2000).

As is typical with studies like these (e.g., Kenney, Silver, Alacaci, & Zawojewski, 1998), the raters found the cognitive demands category to be the most difficult to code, partly because items can often be solved in multiple ways, sometimes as a function of the examinee's proficiency. What might be a problem-solving item for one examinee might require another to apply extensive procedural knowledge. For instance, consider an item asking students for the sum of the first 101 numbers starting with zero. A procedural knowledge approach might involve a computation-intensive method, such as entering all the numbers into a calculator to obtain the resulting sum. However, a problem-solving approach would entail a recognition that all the numbers, except the number 50, can be paired with another number to form a sum of 100 (100+0, 99+1, 98+2, etc.). The total sum is then computed by multiplying the number of pairs (i.e., 50) by 100 and adding 50.

Although distinctions among different levels of cognitive process cannot always be separated neatly, "what can be classified are the actions a student is likely to undertake in processing information and providing a satisfactory response. Thus, … assessment tasks [can be] classified according to the categories they most closely represent in terms of the type of processing they might be expected to require" (NAGB, 2000, p.1 of Chapter 4). In the example above, a procedural knowledge approach may yield the correct answer, but it is unlikely that the solution can be obtained in a reasonable amount of time. Instead, a problem-solving approach is more likely to be used to solve the question. Consequently, for items that can be solved in multiple ways, raters coded for

the cognitive process that was most likely to be elicited. Judgments about the cognitive processes most likely to be evoked were based on raters' prior experience in which they observed high school students as the students solved math problems.

ELA Coding Categories

Coding categories in ELA cover three types of skills: reading, editing, and writing. Reading skills relate to students' vocabulary and comprehension of reading passages. Editing skills relate to students' ability to recognize sentences that violate standard written English conventions. Writing skills pertain to how well students can produce a composition that clearly and logically expresses their ideas.[2] Many of the tests include two or all three of these skills, but some assessments focus on a single type, namely writing. Because many of the tests measuring reading or editing skills include reading passages followed by sets of items, it was necessary to categorize both the reading passage and the individual items.

As with math, the ELA coding categories summarize the technical features, content, and cognitive demands of each assessment. The technical features category in ELA describes time limit, test length, and format. There are only two possible formats for ELA assessments, multiple-choice or open-ended. In ELA, open-ended items require students to produce a writing sample.

The content category is different from that in math because content areas are not as clearly delineated in English. Whereas a math item may be classified into specific content areas such as elementary algebra, geometry, or so forth, there are not analogous English content areas in which to classify reading passages. Instead, the content category in English describes the subject matter of the reading passage (topic), the author's writing style (voice), and the type of reading passage (genre). In other words, there are three dimensions to the ELA content category: topic, voice, and genre. Topic consists of five levels—fiction, humanities, natural science, social science, and personal accounts. Voice consists of four levels—narrative, descriptive, persuasive, and informative. Genre also contains four levels—letter, essay, poem, and story.

[2] Some reading measures require examinees to produce a writing sample that demonstrates comprehension of a reading passage. In this case, the test is considered a measure of both reading and writing skills, although it is understood that measuring reading skills is the test's primary purpose.

For exams measuring reading or editing skills, raters used all three dimensions of the content category to describe the reading passages. However, all three dimensions are not relevant to writing tests, as such tests do not include reading passages. Instead, writing tests contain a short prompt that introduces a topic that serves as the focus of students' compositions. Because students are not instructed to use a particular voice or genre for their compositions, only the topic dimension is needed to describe the writing prompts. Table 2.2 provides more details of the content category used for ELA assessments.

Table 2.2
Descriptions of the ELA Content Category

Dimension	Description or Example	Used for Reading Skills	Used for Editing Skills	Used for Writing Skills
Topic		Y	Y	Y
Fiction	Writing based on imaginary events or people			
Humanities	e.g., artwork of Vincent Van Gogh			
Natural sciences	e.g., the reproductive process of fish			
Social sciences	e.g., one man, one vote; cost effectiveness of heart transplants			
Personal accounts	e.g., diary account of death of a parent			
Voice		Y	Y	N
Narrative	Stories, personal accounts, personal anecdotes			
Descriptive	Describes person, place, or thing			
Persuasive	Attempt to influence others to take some action or to influence someone's attitudes or ideas			
Informative	Share knowledge; convey messages, provide information on a topic, instructions for performing a task			
Genre		Y	Y	N
Letters				
Essays				
Poems				
Stories				

For measures of reading and editing skills, the cognitive demands category is intended to capture different levels of cognitive processes, ranging from low to high levels. Using similar coding categories as those in previous alignment studies (Education Trust, 1999), we distinguished among three levels of cognitive processes: recall, evaluate style, and inference. Table 2.3 provides descriptions of each of these levels.

In reading, questions that could be answered via direct reference to the passage are recall items, whereas questions that require examinees to interpret the material are inference items. Questions that pertain to the development of ideas or improve upon the presentation of the reading passages are coded as evaluating style. For editing measures, items that entail application of grammatical rules are recall items. Typically, most of these questions concern mechanics or usage errors. Inference items are those that require examinees to identify cause-and-effect relationships, and evaluating style items relate to rhetoric skills, such as sentence organization or clarity.

Table 2.3
Descriptions of the ELA Cognitive Demands Category for Tests Measuring Reading or Editing Skills

Cognitive Process	Description or Example
Recall	Answer can be found directly in the text, or by using the definitions of words or literary devices, or by applying grammatical rules
Infer	Interpret what is already written
Evaluate style	Improve the way the material is written

The above cognitive demands category is not applicable to writing measures because students do not respond to items, but instead write their own compositions. For writing tests, the cognitive demands category focuses on the scoring criteria, particularly the emphasis given to mechanics, word choice, organization, style, and insight. The descriptions of these elements are provided in Table 2.4.

Table 2.4
Description of the ELA Cognitive Demands Category for Tests Measuring Writing Skills

Scoring Criteria	Description or Example
Mechanics	Grammar, punctuation, capitalization
Word choice	Use of language, vocabulary, sentence structure
Organization	Logical presentation, development of ideas, use of appropriate supporting examples
Style	Voice, attention to audience
Insight	Analytic proficiency, accurate understanding of stimulus passage, thoughtful perceptions about its ramifications

Rater Agreement in Applying the Coding Categories

Two raters, who had expertise in the relevant subject area, judged alignment by applying the coding categories to each item. One rater coded the math assessments, and the other rater coded the English assessments. An additional rater coded eight tests, 4 each in math and English. All discrepancies were resolved through discussion. Consensus was fairly high, as kappa statistics ranged from .80 to 1.0 (i.e., perfect agreement) for all but two categories.[3] (For specific percent agreement for each coding category, see Appendix B). One of the exceptions was the content category in math, where items often assessed skills in more than one area. Agreement in this category was .76. The final exception was the cognitive demands category in math, where kappa values ranged from .42 to .63, with an average of .56.

Evaluating the Extent of Alignment Among Tests

In interpreting the results, an important issue concerns the standard with which to evaluate the extent of alignment. Specifically, how large should the discrepancies be before we consider two tests to be poorly aligned? To guide our decisions, we analyzed data from an alignment study conducted by Education Trust (1999). We found the average discrepancy across coding categories to be approximately 24%, with a standard deviation of 19%.[4] That is, differences between any two tests were typically within the

[3] Kappa values are measures of agreement after correcting for chance level of agreement.
[4] To calculate the average discrepancy, we found the absolute value of the difference between each pair of tests for each coding category, and averaged these values.

range of 5%-43%. Using these results as a guideline, we decided that differences of 25% or less are considered "small," (i.e., the tests are well-aligned), differences between 25% and 50% are "moderate," (i.e., the tests are moderately aligned), and differences greater than 50% are "large" (i.e., the tests are not well-aligned). Thus, in order to be a misalignment, discrepancies between tests must be greater than 50% *and* cannot be attributed to differences in intended test use.[5]

As there is currently no consensus among researchers regarding the standards for judging the extent of alignment among assessments, the above criteria should not be interpreted as a definitive standard. Additionally, because there has been no research to date that has established how large discrepancies among tests must be before they send students received mixed signals regarding the skills needed to be prepared for college-level work, differences that are considered "small" can nevertheless have important implications. Thus, categorizations of discrepancies as "small," "moderate," or "large" should be interpreted cautiously, and the study's focus should be viewed as mainly descriptive and comparative.

Limitations

Although the use of expert judgments is a fairly common approach to studying alignment (e.g., Kenney, Silver, Alacaci, Zawojewski, 1998; Webb, 1999), this study does not provide a complete picture of these assessments. We have not, for example, systematically examined differences in content standards or test specifications, which may account for some of the discrepancies among exams. Furthermore, an analysis of scores might reveal that seemingly different instruments rank order or classify students roughly equivalently. Similarly, observations and interviews with students as they take the tests, an approach that is sometimes used during the test development process, could result in somewhat different interpretations of a test's reasoning requirements.

Results were similar for math and English. The average discrepancy in math was approximately 24% with a standard deviation of 19%; in English, the average discrepancy was approximately 23% with a standard deviation of 18%.

[5] Readers are reminded that the content category in math and the cognitive demands category in both subjects are the focus of this paper. Thus, large discrepancies with respect to the technical features category or to the ELA content category will not be considered misalignments.

Finally, increasing the number of forms studied for each assessment may enhance the generalizability of our findings. Although we attempted to examine all available forms, this was not always possible.[6] Namely, college admissions measures and some commercially-available college placement exams have multiple, parallel versions. Despite the fact that parallel forms are intended to have similar content and structure, tests represent a sample of skills from a single testing occasion, so forms from other occasions will vary to some extent. This is especially true when we analyze alignment among English Language Arts (ELA) topics, where any given test form provides a limited sample (e.g., there may be only one reading passage).

[6] We were able to examine all versions of the college placement measures used by the selective institutions. Additionally, we were able to examine all versions of the state achievement tests, except California's Stanford 9. For these two categories of tests, our sample is large relative to the entire domain.