National Assessment Governing Board

Committee on Standards, Design and Methodology

August 7, 2015
10:15 am – 12:30 pm

AGENDA

10:15 – 10:20 am	Introductions and Review of Agenda *Lou Fabrizio, COSDAM Chair*	
10:20 – 10:55 am	Content Alignment Studies of Grade 8 NAEP and ACT Explore in Reading and Mathematics *Rolf Blank, NORC at the University of Chicago*	Attachment A
10:55 – 11:35 am	Statistical Linking Studies of Grade 8 NAEP and ACT EXPLORE in Reading and Mathematics *Andreas Oranje, Educational Testing Service*	Attachment B
11:35 – 11:40 am	Other Issues and Questions *COSDAM Members*	
11:40 – 11:45 am	BREAK	
11:45 am – 12:30 pm	**CLOSED SESSION:** Project Update for Technology and Engineering Literacy Achievement Levels Setting *Kelly Burling, Pearson*	Secure materials sent under separate cover
	Information Item • Update on Evaluation of NAEP Achievement Levels • NAEP Job Training Preparedness Report • Procurement on Participation in NAEP: Critical Review and Synthesis of Research	Attachment C Attachment D Attachment E

NAEP-ACT EXPLORE Content Alignment Studies
Project Results

Contract ED-NAG-14-C-0002

In September 2014, NORC at the University of Chicago, along with its subcontractor, the Wisconsin Center for Education Products and Services (WCEPS), were awarded a contract to conduct content alignment studies with the ACT EXPLORE assessments in reading and mathematics and the 2013 National Assessment of Educational Progress (NAEP) Reading and Mathematics assessments at grade 8. The purpose of this research is to evaluate the extent to which 8th grade NAEP is aligned in content and complexity with the EXPLORE assessment. For each subject area, the studies compared the two assessments (NAEP and ACT EXPLORE) to the NAEP frameworks, and also to the ACT College Readiness Standards. The results of these NAEP-EXPLORE content comparisons will also inform interpretations from statistical linking studies of 2013 results of NAEP and EXPLORE in grade 8 reading and mathematics.

To support the provision of ACT proprietary EXPLORE data, the Governing Board also issued a sole source contract with ACT, Inc. NORC worked with ACT to receive data and materials that were used in the content alignment studies, and consulted with ACT assessment staff to support the work and analyses.

Study Design

The content alignment methodology is based on a design produced by Norman Webb in 2009, commissioned by the Governing Board as part of the 12th grade preparedness research program. One key feature of the specified design for analyzing the alignment between the NAEP mathematics and reading assessments and the ACT EXPLORE assessments was to conduct a framework analysis comparing the two frameworks for the assessments. The purpose of the framework analysis was to determine the extent to which the documents that are intended to specify the domain of knowledge to be assessed are the same or different. A second feature of the study design was to conduct a Content Alignment Institute (CAI) that is structured around panels of content experts, including teachers, who map the items from each assessment to each of the content frameworks. The alignment between the two assessments is determined by comparing the mapping of both assessments to each of the two frameworks. The alignment between these two assessments will be gauged by the extent of overlapping content knowledge targeted by the two assessments and by the extent of content knowledge that is targeted and unique for each assessment.

Project leaders at NORC and WCEPS successful implemented the CAI in February 2015. Thirty-two panelists (16 math experts, 16 reading experts), four facilitators (two math, two reading), and representatives from NCES, ACT and NAGB comprised the participants at the

Institute held the NORC Bethesda, MD, facility during that week. A national process of outreach and recruitment was conducted by NORC to ensure that panels would have members who are experienced, qualified teachers and assessment specialists in reading and mathematics, and that the panels would be representative of students and teachers based on gender, race/ethnicity, and region of the U.S.

In the CAI, the content analysis of reading and mathematics assessments was conducted by two panels of eight educators for each content area. A panel of eight constitutes a sufficient number to insure high reliability of the assigned depth-of-knowledge level to a standard or assessment item and the reliability of the assigned assessment item to a content standard. Two panels were included in the design to identify and analyze potential variations in coding results that may reflect legitimate differences. Another feature incorporated into the design for collecting these data is the adjudication of coding results. In adjudication, panelists discuss their differences in their initial coding results to determine the degree of variation in coding among the group, after panelists have initially reviewed and individually mapped items on an assessment to the standards and objectives in the framework. Facilitators were trained to guide the discussion to help panelists identify agreement.

Outcomes and Results

The implementation of the NAEP-EXPLORE content alignment study followed very closely to the specified design. There were only a few deviations. There were time pressures to complete all of the work at the five-day institute. The study reports for mathematics and reading are still being finalized, incorporating reviews by NAGB and ACT staffs. Rolf Blank of NORC at the University of Chicago will present highlights from the study findings to COSDAM. The study reports will be posted to the Governing Board academic preparedness website during the late summer or early fall.

NAEP Academic Preparedness Research

Update on State Statistical Linking Studies with ACT EXPLORE®

In this presentation, we will update the Committee on Standards, Design, and Methodology (COSDAM) on the most recent statistical linking work, which is part of a second phase of academic preparedness research. The first phase of the National Assessment Governing Board's statistical linking research, part of a broader academic preparedness research agenda, was based on 2009 data and included a national NAEP-SAT linking as well as in-depth linking and analysis of Florida's longitudinal database. The second phase is based on 2013 data and includes several statistical linking studies at the state level. One particular interest is to investigate the extent to which 8th graders are on track for being academically prepared for college once they reach the end of high school. To that end, statistical linking studies between 8th grade NAEP (Reading and Mathematics) and EXPLORE®, a test[1] developed and administered by ACT, Inc., were conducted. The EXPLORE® assessment is linked to performance on the ACT, and on-track preparedness benchmarks have been established. The study was conducted in three states (Kentucky, North Carolina, and Tennessee), where EXPLORE® was administered to all students state-wide who were in grade 8 during the 2012-13 school year. For students participating in NAEP, their EXPLORE® scores were provided by the states (via data sharing agreements) and linked, using a process that protects student confidentiality.

The grade 8 state-level statistical linking studies were designed to pursue three specific analysis questions that guided methodological choices for linking and validation:

1) What are the correlations between grade 8 NAEP and EXPLORE® scores in Reading and Mathematics?
2) What scores on the grade 8 NAEP Reading and Mathematics scales correspond to the EXPLORE® college readiness benchmarks?
3) What are the average grade 8 NAEP Reading and Mathematics scores and the inter-quartile ranges (IQR) for students below, at, and above the EXPLORE® college readiness benchmarks?

In this session, research findings from the state statistical linking studies will be presented to COSDAM. The correlations between NAEP and EXPLORE® Reading and Mathematics scale scores were not sufficiently strong enough to support concordance for any of the states, and, therefore, statistical projection was applied to characterize the relationship between the two assessment scales. We will show benchmark estimates as well as resultant preparedness percentages. Finally, we will provide an overview of next steps in terms of other statistical linking studies that are currently planned or underway.

[1] ACT will discontinue the use of the EXPLORE® test after fall 2015 for existing users and no new users are now being accepted.

Discussion Draft

NAEP Grade 8 Academic Preparedness Research:

Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Kentucky Students

Adrienne Sgammato
Mei-Jang Lin
Laura Jerry
David Freund
Rochelle Michel
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
NAGB Staff: Sharyn Rosenberg, Assistant Director for Psychometrics

Prepared by Educational Testing Service for the National Center of Education Statistics under contract ED-IES-13-C-0017, Task 9, Option 9(C) at the request of the National Assessment Governing Board.

Introduction

Starting in early 2003, the National Assessment Governing Board embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on *academic preparedness*. To be "academically prepared for college", 12th graders should have the knowledge and skills in reading and mathematics to qualify for placement into entry-level, credit-bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research (National Assessment Governing Board, 2009), potential benchmarks were identified on the 12th grade Reading and Mathematics scales to indicate what level of performance would correspond to a reasonable probability of being academically prepared for postsecondary education. As a result, a national postsecondary education preparedness percentage could be estimated and reported for the 2013 assessments in Reading and Mathematics. Details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (http://www.nagb.gov/what-we-do/commission.html).

As part of the initial statistical linking research, Florida participated (and continues to participate) at the 12th grade level and was a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

Moving forward, one focus of the second phase of the NAEP academic preparedness research is to study the extent to which grade 8 students are on track for being academically prepared for college by the end of high school. Several states, including Kentucky, participated in the statistical linking research and provided data on students who were part of the NAEP grade 8 sample during the 2012-2013 school year. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.

In this report we will describe NAEP and EXPLORE® assessments in Reading and Mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

Linking Assessments

The ACT EXPLORE® Assessment

The EXPLORE® test[2] developed by ACT was administered to nearly all 8th graders in Kentucky during the 2012-2013 school year (with the testing window from Sep 17 to Sep 28, 2012). The assessment includes four multiple-choice tests. Each test measures student's achievement in one of the following four areas: English, Mathematics, Reading, and Science. Students had 30 minutes to finish each test. The number of items in the test varies by subject, for reading and mathematics, both tests have 30 items. EXPLORE® scores provide evidence about the knowledge and skills that students are likely to have in each of the four aforementioned areas. The distribution of item difficulties was selected so that the tests will effectively differentiate among students who vary widely in the level of achievement. A composite score is provided, which is calculated as the average of the four test scores. The individual test scores, as well as the composite score, range from 1 to 25 and are disseminated to students and schools directly. In this study, only the Reading and Mathematics scores were used to link with the NAEP Reading and Mathematics assessments.

The ACT EXPLORE® assessments were designed to assess a specific student's academic progress at the 8th and 9th grade levels, especially with respect to college and career readiness. To help students translate test scores into a clear indicator of their current level of college readiness, ACT derived the ACT College Readiness Benchmarks based on a review of normative data, college admissions criteria, and information obtained through ACT's Course Placement Services. Students who meet a benchmark on the ACT test have approximately a 50% chance of obtaining a B or higher and approximately a 75% chance of obtaining a C or higher in the corresponding credit-bearing first-year college courses (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). In addition, there are corresponding benchmarks for the ACT EXPLORE®, which are linked to the ACT College Readiness Benchmarks. Students who meet a benchmark on the EXPLORE® test have approximately a 50% chance of meeting the ACT Benchmark in the same subject, and are likely to have approximately the same chance of earning a B or better grade in the corresponding college course(s) by the time they graduate high school. The current College Readiness Benchmarks for the EXPLORE® Reading test for grade 8 is 16 and for the EXPLORE® Mathematics test is 17 (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). These benchmarks were used in this investigation. Note that the EXPLORE® reading benchmark was adjusted in 2013. Previously the reading EXPLORE® benchmark was 15. The math EXPLORE® benchmark remained unchanged.

[2] ACT will discontinue the use of the EXPLORE® test after fall 2015 for existing users and no new users are now being accepted.

The National Assessment of Educational Progress (NAEP)

The NAEP test was administered to selected 8th graders in Kentucky during the 2012-2013 school year (with the testing window from the last week of Jan to the first week of Mar in 2013). NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as Reading, Mathematics, and Science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame in order to produce results that are nationally representative and also representative of participating states and urban districts. Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 8 reading, each booklet contains two blocks of about 10 items each. For grade 8 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score *can* be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods. For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy, Beaton, Kaplan, & Sheehan (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., EXPLORE® scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 8 Reading and Mathematics and the goal is to express the aforementioned ACT EXPLORE® benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one assessment subject. Consequently, each student in the matched or linking sample had EXPLORE® scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

Linking

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong and the instruments have a similarly high reliability, we would be able to claim that the two scales are

largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a 'best' projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., "at score level X, students have a reasonably high probability to be prepared"). In the case of the preparedness linking studies, and taking past studies into account, a moderate relationship is most probable. We will elaborate further on this in subsequent sections.

Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. Content alignment studies between NAEP and EXPLORE® Reading and Mathematics are being conducted by the National Opinion Research Center (NORC) at the University of Chicago (under contract ED-NAG-14-C-0002 with the National Assessment Governing Board) and will provide an important context for the statistical linking results presented here.

Methodology

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

Data

This study used data from students who were sampled and assessed in NAEP 8th grade reading or mathematics in 2013 and had also taken the EXPLORE® assessment. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered to samples of 8th grade students that were representative of each state, and together of the nation. As a result, about 2,700 public school students were sampled from each state for each subject. In addition to state representative samples, NAEP also assesses many large urban districts including Jefferson County in Kentucky, adding a representative sample for those districts. Consequently, about 3,700 and 3,800 students were assessed in reading and mathematics, respectively, in Kentucky. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The EXPLORE® assessment is required in Kentucky at the 8th grade level, meaning that almost all students who were sampled for NAEP also participated in EXPLORE® and have associated scores. The reverse is obviously not true, given that NAEP is sample-based (i.e., not every student who participated in EXPLORE® also participated in NAEP).

The process of matching EXPLORE® scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics (NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work. In addition, data confidentiality agreements were established between all parties involved and the Kentucky Department of Education. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of state supplied scores (e.g., EXPLORE® scores) was assured through the assignment of a pseudo ID for students taking that assessment and using that pseudo ID as a way to transfer scores to ETS *without* the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again *without* any PII. Via the pseudo ID, ETS subsequently matched EXPLORE® scores to NAEP files. In the case of Kentucky, EXPLORE® scores were matched at 96%, which is extraordinarily high. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) were at or above 92%. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates.

Table 1. Weighted percentages by gender and race of the Kentucky linking samples

Reading								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	43%	5%	2%	1%	#[1]	#	1%	**51%**
Female	41%	5%	2%	1%	#	#	1%	**49%**
***Total*[2]**	**83%**	**10%**	**4%**	**1%**	**#**	**#**	**2%**	**100%**
							Overall Match Rate	**96%**
Mathematics								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	42%	5%	2%	1%	#	#	1%	**51%**
Female	41%	5%	2%	1%	#	#	1%	**49%**
***Total*[2]**	**83%**	**10%**	**4%**	**1%**	**#**	**#**	**2%**	**100%**
							Overall Match Rate	**96%**

NOTES: [1]# Rounds to zero.
[2] Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have very different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a *higher* stakes test (i.e., EXPLORE® test) and very low on the *lower* stakes test, the low

performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered 'outlier' and removed from further analyses. An initial examination of the joint distribution of NAEP and EXPLORE® revealed very few potential outlier cases. After this more cursory inspection, standardized residuals from robust regression (Huber, 1973) were used to identify approximately 0.4% of cases in both reading and mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). We refer to Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded from the final linking samples and were not used in subsequent analyses.

Analysis Approach

After preparatory data identification, matching, merging, and data reconciliation, the linking analyses were conducted. The current study was designed to pursue three specific analysis questions that guide the choices in methodology for the linking and validation:

1) What are the correlations between the grade 8 NAEP and EXPLORE® scores in reading and mathematics?
2) What scores on the grade 8 NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks?
3) What are the average grade 8 NAEP reading and mathematics scores (and the difference between the 75th and 25th percentiles) and the IQR for students below, at, and above the EXPLORE® benchmarks?

We will describe pertinent methodological details about the analyses followed by the results of the analyses in the final section. The key steps of the analyses are (a) estimating the correlation between NAEP and EXPLORE®, which includes use of the aforementioned latent regression methodology (b) determining the appropriate methodology for linking based on those correlations (c) applying procedures to effectively estimate cumulative probability functions and (d) calculating impact data as part of the results.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is that NAEP measures constructs that are represented on item response theory based latent scales, which are not measured reliably at the student level. However, pertinent data from students in specified groups of interest can be pooled to estimate reliable scores at the group level. EXPLORE® scores, on the other hand, are reliably estimated at the individual level and can be treated as a set of consecutive (semi-continuous) groups. Correlations between NAEP and EXPLORE® can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is 0.74 and for mathematics is 0.81. While these are not low correlations, they do suggest that there is enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score *x* on NAEP exactly corresponds to a score *y* on EXPLORE® and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of *x* on NAEP corresponds to the proportion *p* of students attaining the benchmark score of *y* or higher on EXPLORE®. Subsequently, a choice for *p* has to be made, where a more conservative claim requires a higher *p*. This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high *p* has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and EXPLORE® reading (r =0.74) and mathematics (r = 0.81) are not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is r = 0.866 (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection was used in this study. As mentioned before, typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-EXPLORE® distributions[3].

An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size[4] in order to produce reliable comparisons. For the Kentucky linking samples, both gender groups met

[3] For reading, as part of the loglinear smoothing procedure we preserved the first 2 moments for the NAEP distribution, 4 moments for the EXPLORE® distribution, and 4 cross-moments. For math, we preserved the first 3 moments for the NAEP distribution, 6 moments for the EXPLORE® distribution, and 4 cross-moments. These loglinear smoothing models mostly resulted in the smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006), although model complexity and sample size was also taken into consideration.

[4] The minimum was set at 500 as a rule of thumb, but based on the idea that there is at least one observation below -3 and above +3 standard deviations (in a standard normal distribution) in expectation.

that criterion. For the race/ethnicity groups, only White and Black student subgroups met the criterion. Separate linking functions were established for these subgroups and deviations from the overall linking function indicated violation of invariance. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. For NAEP mathematics, no substantial deviation from the overall linking function was detected for Male, Female, or White student subgroups. The linking function for Black students was slightly lower than the overall linking function. For NAEP reading, the linking function for White students was very close to the overall linking function. But there was some variation in the linking results observed for Male, Female, and Black student subgroups. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projections tables containing conditional cumulative distributions of NAEP proficiencies for EXPLORE® scores. The range of possible NAEP scores below, at, and above the EXPLORE® benchmark (16 on the EXPLORE® reading scale and 17 on the EXPLORE® mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the EXPLORE® benchmarks.

In the following section we will discuss the results of the linking study, focusing on the second and third analysis questions: What NAEP scores correspond to the EXPLORE® benchmarks and what are the distributional characteristics associated with those benchmarks.

Results

On Track Markers

The most important result, following the second and third analysis questions, is to determine what scores on the NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks. In other words, what would be the 'on track to be prepared' score level on NAEP that corresponds most reasonably to an established 'on track' benchmark.

Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and above the EXPLORE® benchmarks are spread

out, though more so for students below the benchmark than above. Note that the mean *at* the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

Table 2: Descriptive NAEP Statistics for Students Below, At, or Above the EXPLORE® Benchmarks

Subject	EXPLORE® Benchmark	Mean	Percentage[2]	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	255	64%	27	237	274	37
	At	281	8%	22	266	296	30
	Above	301	28%	24	284	317	33
Mathematics	*Below*	267	65%	27	250	285	35
	At	296	11%	19	283	308	25
	Above	319	24%	23	302	334	32

NOTES: [1]IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2]Detail may not sum to totals because of rounding.

To determine the NAEP scale score point that most reasonably corresponds to the EXPLORE® benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the EXPLORE® benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned before, a proportion level has to be chosen commensurate the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows at which point students are more likely to have reached the benchmark than not (i.e., the probability is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be 'on track to preparedness'. For context, a secondary, lighter red line indicates when the probability is set at 0.80, indicating a relatively high level of confidence that students have attained the EXPLORE® benchmark.

From the graphs it can be deducted that the location where students have a reasonable probability to be on track for reading could be set at a NAEP scale score of 286, slightly above the *Proficient* achievement level. The mathematics counterpart could be set at 299, which coincides with the *Proficient* achievement level.

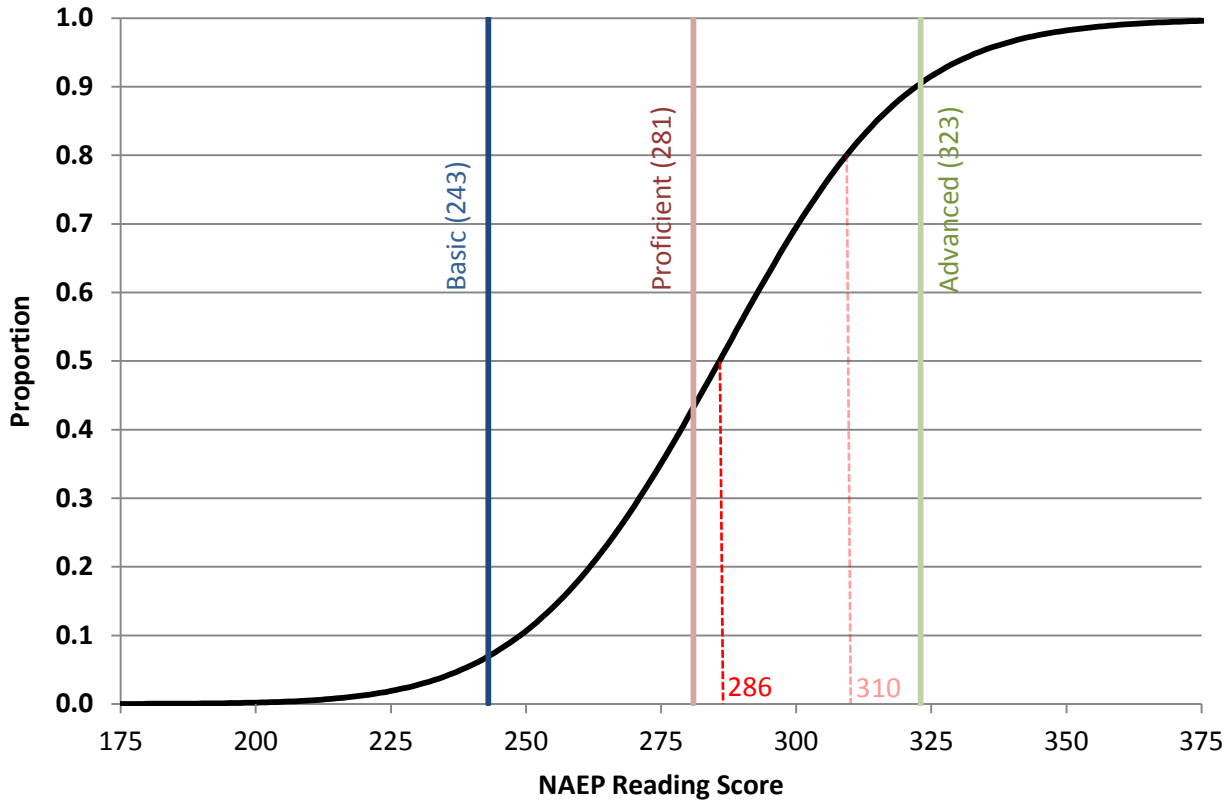

Figure 1: Proportion of students meeting the Reading EXPLORE® benchmark of 16 in Kentucky for NAEP Reading levels

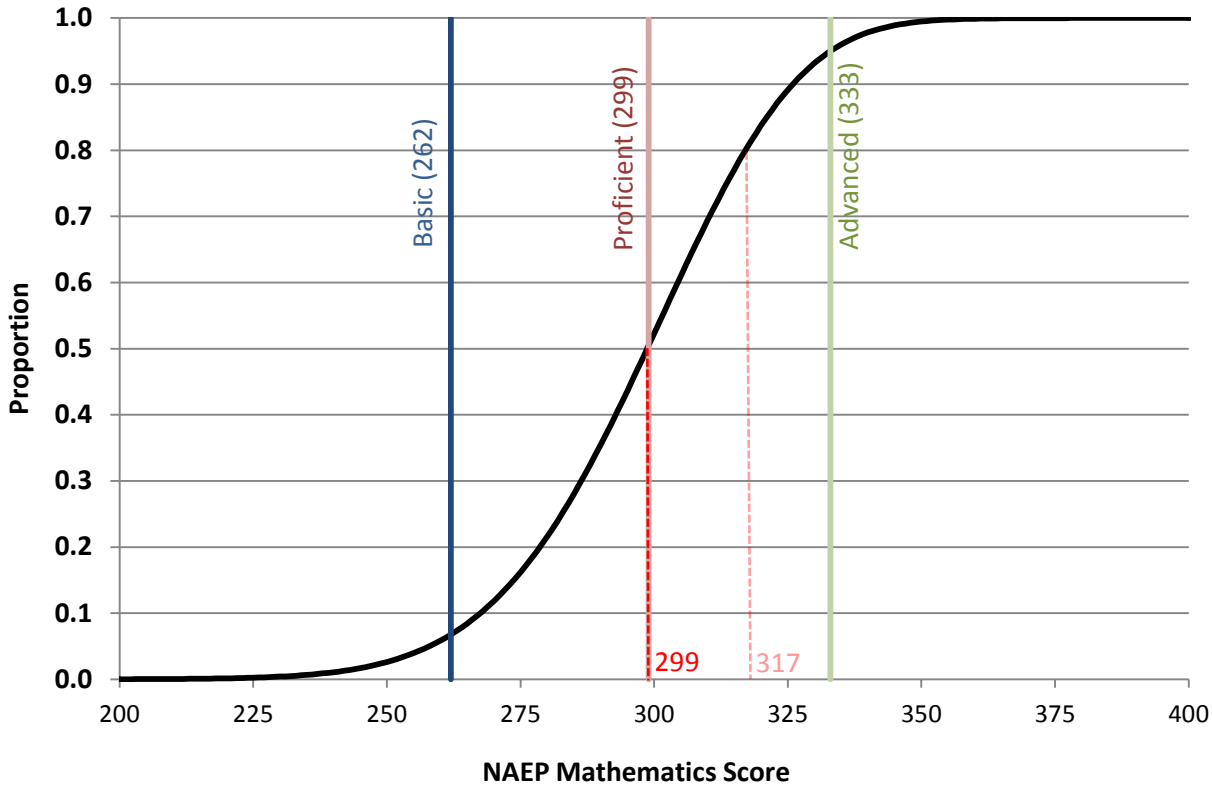

Figure 2: Proportion of students meeting the Mathematics EXPLORE® benchmark of 17 in Kentucky for NAEP Mathematics levels

Impact

Now that potential points have been identified, it is important to show what percentage of students in Kentucky are deemed to have a reasonable probability (i.e., the probability set at 0.50) of being on track in grade 8 across various student groups. Table 3 provides those percentages, based on the potential points identified on the NAEP scales, as well as the ACT EXPLORE® benchmarks. Table 3 indicates that overall about 31 to 35 percent of students are on track, but the results differ across different subgroups. No significance testing has been conducted to compare these percentages and, therefore, no comparative statements will be made.

Table 3: Percentage of the Kentucky linking samples that have a reasonable probability to be on track to be academically prepared based on the potential points identified on the NAEP scale, compared to the percentage of the same sample meeting the Reading EXPLORE® benchmark of 16 and Mathematics EXPLORE® benchmark of 17.

	Reading		Mathematics	
Student Group	**NAEP ≥ 286**	**EXPLORE® ≥ 16**	**NAEP ≥ 299**	**EXPLORE® ≥ 17**
Total	32%	34%	31%	35%
Male	28%	29%	32%	36%
Female	36%	39%	31%	35%
White	34%	37%	34%	38%
Black	12%	16%	12%	15%
Hispanic	30%	25%	19%	25%

Summary

The goal of this study was to statistically relate NAEP and EXPLORE® and use that relationship to identify a reference point or range on the NAEP 8th grade reading and mathematics scales reasonably associated with ACT's preparedness benchmarks on the EXPLORE® reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 8th grade who are on track to be prepared for college for the nation and for states. The first step involves three participating states, including Kentucky, who have graciously provided the critical EXPLORE® data necessary to calculate the relationship with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for 'on track to preparedness' reporting (see Figures 1 and 2 for examples of how the markers were determined).

A key finding was that the relationship between the two scales is moderate, meaning that the kind of relational statements that can be made need to be presented in notions of probability rather than direct one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct, however, it does make interpretation somewhat more challenging. The results showed that NAEP scale score points at or just above the *Proficient* achievement levels could form a reasonable basis for reporting 'on track for preparedness'. Approximately 32% of Kentucky 8th graders met that criterion for reading and 31% met the criterion for math. Further content alignment work, which is conducted independently from this study, should provide further context to these results.

References

ACT EXPLORE Technical Manual 2013/2014. (http://www.act.org/explore/pdf/TechManual.pdf)

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*.

Discussion Draft

NAEP Grade 8 Academic Preparedness Research:

Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for North Carolina Students

Adrienne Sgammato
Mei-Jang Lin
Laura Jerry
David Freund
Rochelle Michel
Nuo Xi
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
NAGB Staff: Sharyn Rosenberg, Assistant Director for Psychometrics

Prepared by Educational Testing Service for the National Center of Education Statistics under contract ED-IES-13-C-0017, Task 9, Option 9(C) at the request of the National Assessment Governing Board.

Introduction

Starting in early 2003, the National Assessment Governing Board embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on *academic preparedness*. To be "academically prepared for college", 12th graders should have the knowledge and skills in reading and mathematics to qualify for placement into entry-level, credit-bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research (National Assessment Governing Board, 2009), potential benchmarks were identified on the 12th grade Reading and Mathematics scales to indicate what level of performance would correspond to a reasonable probability of being academically prepared for postsecondary education. As a result, a national postsecondary education preparedness percentage could be estimated and reported for the 2013 assessments in Reading and Mathematics. Details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (http://www.nagb.gov/what-we-do/commission.html).

As part of the initial statistical linking research, Florida participated (and continues to participate) at the 12th grade level and was a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

Moving forward, one focus of the second phase of the NAEP academic preparedness research is to study the extent to which grade 8 students are on track for being academically prepared for college by the end of high school. Several states, including North Carolina, participated in the statistical linking research and provided data on students who were part of the NAEP grade 8 sample during the 2012-2013 school year. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.

In this report we will describe the NAEP and EXPLORE® assessments in Reading and Mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

Linking Assessments

The ACT EXPLORE® Assessment

The EXPLORE® test[5] developed by ACT was administered to nearly all 8th graders in North Carolina during the 2012-2013 school year (with the testing window from Oct 1 to Oct 31, 2012). The assessment includes four multiple-choice tests. Each test measures student's achievement in one of the following four areas: English, Mathematics, Reading, and Science. Students had 30 minutes to finish each test. The number of items in the test varies by subject, for reading and mathematics, both tests have 30 items. EXPLORE® scores provide evidence about the knowledge and skills that students are likely to have in each of the four aforementioned areas. The distribution of item difficulties was selected so that the tests will effectively differentiate among students who vary widely in the level of achievement. A composite score is provided, which is calculated as the average of the four test scores. The individual test scores, as well as the composite score, range from 1 to 25 and are disseminated to students and schools directly. In this study, only the Reading and Mathematics scores were used to link with the NAEP Reading and Mathematics assessments.

The ACT EXPLORE® assessments were designed to assess a specific student's academic progress at the 8th and 9th grade levels, especially with respect to college and career readiness. To help students translate test scores into a clear indicator of their current level of college readiness, ACT derived the ACT College Readiness Benchmarks based on a review of normative data, college admissions criteria, and information obtained through ACT's Course Placement Services. Students who meet a benchmark on the ACT test have approximately a 50% chance of obtaining a B or higher and approximately a 75% chance of obtaining a C or higher in the corresponding credit-bearing first-year college courses (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). In addition, there are corresponding benchmarks for the ACT EXPLORE®, which are linked to the ACT College Readiness Benchmarks. Students who meet a benchmark on the EXPLORE® test have approximately a 50% chance of meeting the ACT Benchmark in the same subject, and are likely to have approximately the same chance of earning a B or better grade in the corresponding college course(s) by the time they graduate high school. The current College Readiness Benchmarks for the EXPLORE® Reading test for grade 8 is 16 and for the EXPLORE® Mathematics test is 17 (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). These benchmarks were used in this investigation. Note that the EXPLORE® reading benchmark was adjusted in 2013. Previously the reading EXPLORE® benchmark was 15. The math EXPLORE® benchmark remained unchanged.

[5] ACT will discontinue the use of the EXPLORE® test after fall 2015 for existing users and no new users are now being accepted.

The National Assessment of Educational Progress (NAEP)

The NAEP test was administered to selected 8th graders in North Carolina during the 2012-2013 school year (with the testing window from the last week of Jan to the first week of Mar in 2013). NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as Reading, Mathematics, and Science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame in order to produce results that are nationally representative and also representative of participating states and urban districts. Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 8 reading, each booklet contains two blocks of about 10 items each. For grade 8 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score *can* be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods. For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy, Beaton, Kaplan, & Sheehan (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., EXPLORE® scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 8 Reading and Mathematics and the goal is to express the aforementioned ACT EXPLORE® benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one assessment subject. Consequently, each student in the matched or linking sample had EXPLORE® scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

Linking

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong and

the instruments have a similarly high reliability, we would be able to claim that the two scales are largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a 'best' projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., "at score level X, students have a reasonably high probability to be prepared"). In the case of the preparedness linking studies, and taking past studies into account, a moderate relationship is most probable. We will elaborate further on this in subsequent sections.

Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. Content alignment studies between NAEP and EXPLORE® Reading and Mathematics are being conducted by the National Opinion Research Center (NORC) at the University of Chicago (under contract ED-NAG-14-C-0002 with the National Assessment Governing Board) and will provide an important context for the statistical linking results presented here.

Methodology

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

Data

This study used data from students who were sampled and assessed in NAEP 8th grade reading or mathematics in 2013 and had also taken the EXPLORE® assessment. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered to samples of 8th grade students that were representative of each state, and together of the nation. As a result, about 2,700 public school students were sampled from each state for each subject. In addition to state representative samples, NAEP also assesses many large urban districts including Charlotte-Mecklenburg Schools in North Carolina, adding a representative sample for those districts. Consequently, about 4,000 and 3,900 students were assessed in reading and mathematics, respectively, in North Carolina. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The EXPLORE® assessment is required in North Carolina at the 8th grade level, meaning that almost all students who were sampled for NAEP also participated in EXPLORE® and have associated scores. The reverse is obviously not true, given

that NAEP is sample-based (i.e., not every student who participated in EXPLORE® also participated in NAEP).

The process of matching EXPLORE® scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics (NCES) to have NAEP contractors Westat and ETS conduct the preparedness research work. In addition, data confidentiality agreements were established between all parties involved and the North Carolina Department of Public Instruction. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of state supplied scores (e.g., EXPLORE® scores) was assured through the assignment of a pseudo ID for students taking that assessment and using that pseudo ID as a way to transfer scores to ETS *without* the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again *without* any PII. Via the pseudo ID, ETS subsequently matched EXPLORE® scores to NAEP files. In the case of North Carolina, EXPLORE® scores were matched at 96%, which is extraordinarily high. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) were at or above 92%. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates.

Table 1. Weighted percentages by gender and race of the North Carolina linking samples

Reading								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	28%	14%	6%	1%	#[1]	#	1%	**50%**
Female	26%	14%	7%	1%	#	#	2%	**50%**
***Total*[2]**	**54%**	**28%**	**13%**	**2%**	**1%**	**#**	**3%**	**100%**
							Overall Match Rate	**96%**
Mathematics								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	28%	13%	7%	1%	#	#	1%	**51%**
Female	25%	14%	6%	1%	#	#	1%	**49%**
***Total*[2]**	**54%**	**27%**	**13%**	**2%**	**1%**	**#**	**3%**	**100%**
							Overall Match Rate	**96%**

NOTES: [1]# Rounds to zero.
[2] Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have very different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a *higher* stakes test (i.e., EXPLORE® test) and very low on the *lower* stakes test, the low performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered 'outlier' and removed from further analyses. An initial examination of the joint distribution of NAEP and EXPLORE® revealed very few potential outlier cases. After this more cursory inspection, standardized residuals from robust regression (Huber, 1973) were used to identify approximately 0.6% of cases in reading and approximately 0.8% of cases in mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). We refer to Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded from the final linking samples and were not used in subsequent analyses.

Analysis Approach

After preparatory data identification, matching, merging, and data reconciliation, the linking analyses were conducted. The current study was designed to pursue three specific analysis questions that guide the choices in methodology for the linking and validation:

1) What are the correlations between the grade 8 NAEP and EXPLORE® scores in reading and mathematics?
2) What scores on the grade 8 NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks?
3) What are the average grade 8 NAEP reading and mathematics scores (and the difference between the 75th and 25th percentiles) and the IQR for students below, at, and above the EXPLORE® benchmarks?

We will describe pertinent methodological details about the analyses followed by the results of the analyses in the final section. The key steps of the analyses are (a) estimating the correlation between NAEP and EXPLORE®, which includes use of the aforementioned latent regression methodology (b) determining the appropriate methodology for linking based on those correlations (c) applying procedures to effectively estimate cumulative probability functions and (d) calculating impact data as part of the results.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is that NAEP measures constructs that are represented on item response theory based latent scales, which are not measured reliably at the student level. However, pertinent data from students in specified groups of interest can be pooled to estimate reliable scores at the group level. EXPLORE® scores, on the other hand, are reliably estimated at the individual level and can be treated as a set of consecutive (semi-continuous) groups. Correlations between NAEP and EXPLORE® can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is

0.72 and for mathematics is 0.82. While these are not low correlations, they do suggest that there is enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score *x* on NAEP exactly corresponds to a score *y* on EXPLORE® and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of *x* on NAEP corresponds to the proportion *p* of students attaining the benchmark score of *y* or higher on EXPLORE®. Subsequently, a choice for *p* has to be made, where a more conservative claim requires a higher *p*. This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high *p* has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and EXPLORE® reading (r =0.72) and mathematics (r = 0.82) are not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is r = 0.866 (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection was used in this study. As mentioned before, typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-EXPLORE® distributions[6].

An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student

[6] For reading, as part of the loglinear smoothing procedure we preserved the first 3 moments for the NAEP distribution, 4 moments for the EXPLORE® distribution, and 4 cross-moments. For math, we preserved the first 3 moments for the NAEP distribution, 6 moments for the EXPLORE® distribution, and 4 cross-moments. These loglinear smoothing models mostly resulted in the smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006), although model complexity and sample size was also taken into consideration.

groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size[7] in order to produce reliable comparisons. For the North Carolina linking samples, both gender groups met that criterion. For the race/ethnicity groups, only White, Black, and Hispanic student subgroups met the criterion. Separate linking functions were established for these subgroups and deviations from the overall linking function indicated violation of invariance. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. The comparison results showed some variance across the three ethnicity subgroups for both reading and mathematics. In general, the linking functions for the Black and Hispanic subgroups were lower than the overall linking function, and the linking function for White was slightly higher than the overall linking function. The two gender subgroups did not show substantial variation from the overall linking results. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projections tables containing conditional cumulative distributions of NAEP proficiencies for EXPLORE® scores. The range of possible NAEP scores below, at, and above the EXPLORE® benchmark (16 on the EXPLORE® reading scale and 17 on the EXPLORE® mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the EXPLORE® benchmarks.

In the following section we will discuss the results of the linking study, focusing on the second and third analysis questions: What NAEP scores correspond to the EXPLORE® benchmarks and what are the distributional characteristics associated with those benchmarks.

Results

On Track Markers

The most important result, following the second and third analysis questions, is to determine what scores on the NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks. In other words, what would be the 'on track to be prepared' score level on NAEP that corresponds most reasonably to an established 'on track' benchmark.

[7] The minimum was set at 500 as a rule of thumb, but based on the idea that there is at least one observation below -3 and above +3 standard deviations (in a standard normal distribution) in expectation.

Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and above the EXPLORE® benchmarks are spread out, though more so for students below the benchmark than above. Note that the mean *at* the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

Table 2: Descriptive NAEP Statistics for Students Below, At, or Above the EXPLORE® Benchmarks

Subject	EXPLORE® Benchmark	Mean	Percentage[2]	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	252	66%	28	234	271	37
	At	279	8%	22	263	293	30
	Above	298	27%	24	282	314	32
Mathematics	*Below*	270	62%	27	253	289	36
	At	299	12%	19	286	312	26
	Above	323	27%	22	307	338	31

NOTES: [1]IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2] Detail may not sum to totals because of rounding.

To determine the NAEP scale score point that most reasonably corresponds to the EXPLORE® benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the EXPLORE® benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned before, a proportion level has to be chosen commensurate the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows at which point students are more likely to have reached the benchmark than not (i.e., the probability is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be 'on track to preparedness'. For context, a secondary, lighter red line indicates when the probability is set at 0.80, indicating a relatively high level of confidence that students have attained the EXPLORE® benchmark.

From the graphs it can be deducted that the location where students have a reasonable probability to be on track for reading could be set at a NAEP scale score of 285, slightly above the *Proficient* achievement level. The mathematics counterpart could be set at 301, very slightly above the *Proficient* achievement level.

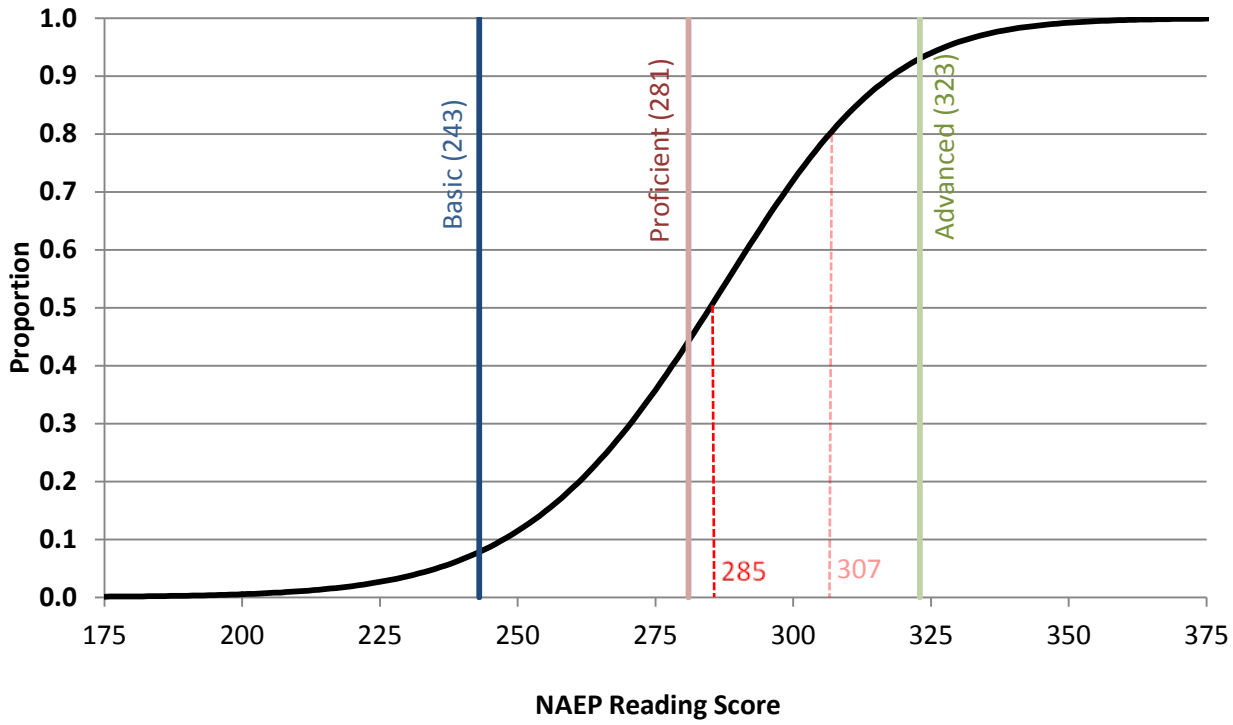

Figure 1: Proportion of students meeting the Reading EXPLORE® benchmark of 16 in North Carolina for NAEP Reading levels

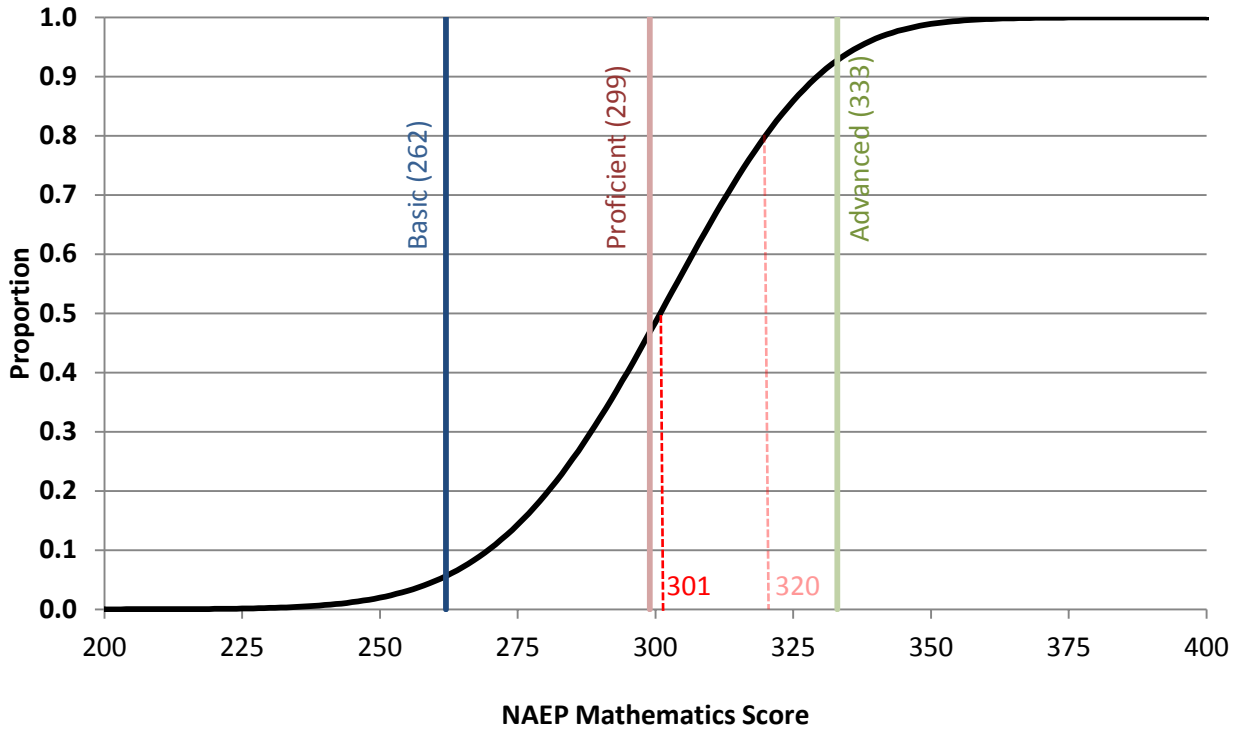

Figure 2: Proportion of students meeting the Mathematics EXPLORE® benchmark of 17 in North Carolina for NAEP Mathematics levels

Impact

Now that potential points have been identified, it is important to show what percentage of students in North Carolina are deemed to have a reasonable probability (i.e., the probability set at 0.50) of being on track in grade 8 across various student groups. Table 3 provides those percentages, based on the potential points identified on the NAEP scales, as well as the EXPLORE® benchmarks. Table 3 indicates that overall about 29 to 38 percent of students are on track, but the results differ across different subgroups. No significance testing has been conducted to compare these percentages and, therefore, no comparative statements will be made.

Table 3: Percentage of the North Carolina linking samples that have a reasonable probability to be on track to be academically prepared based on the potential points identified on the NAEP scale, compared to the percentage of the same sample meeting the Reading EXPLORE® benchmark of 16 and Mathematics EXPLORE® benchmark of 17.

	Reading		Mathematics	
Student Group	**NAEP ≥ 285**	**EXPLORE® ≥ 16**	**NAEP ≥ 301**	**EXPLORE® ≥ 17**
Total	29%	32%	35%	38%
Male	22%	27%	35%	38%
Female	35%	38%	35%	38%
White	39%	42%	46%	49%
Black	14%	17%	16%	18%
Hispanic	19%	22%	27%	28%
Asian	43%	53%	55%	60%

Summary

The goal of this study was to statistically relate NAEP and EXPLORE® and use that relationship to identify a reference point or range on the NAEP 8th grade reading and mathematics scales reasonably associated with ACT's preparedness benchmarks on the EXPLORE® reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 8th grade who are on track to be prepared for college for the nation and for states. The first step involves three participating states, including North Carolina, who have graciously provided the critical EXPLORE® data necessary to calculate the relationship with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for 'on track to preparedness' reporting (see Figures 1 and 2 for examples of how the markers were determined).

A key finding was that the relationship between the two scales is moderate, meaning that the kind of relational statements that can be made need to be presented in notions of probability rather than direct one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct, however, it does make interpretation somewhat more challenging. The results showed that NAEP scale score points just above the *Proficient* achievement levels could form a reasonable basis for reporting 'on track for preparedness'. Approximately 29% of North Carolina 8th graders met that criterion for reading and 35% met the criterion for math. Further content alignment work, which is conducted independently from this study, should provide further context to these results.

References

ACT EXPLORE Technical Manual 2013/2014. (http://www.act.org/explore/pdf/TechManual.pdf)

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*.

Discussion Draft

NAEP Grade 8 Academic Preparedness Research:

Establishing a Statistical Relationship between the NAEP and EXPLORE® Grade 8 Assessments in Reading and Mathematics for Tennessee Students

Adrienne Sgammato
Mei-Jang Lin
Laura Jerry
David Freund
Rochelle Michel
Nuo Xi
Andreas Oranje

NCES Project Officer: Bill Tirre, Senior Technical Advisor
NAGB Staff: Sharyn Rosenberg, Assistant Director for Psychometrics

Prepared by Educational Testing Service for the National Center of Education Statistics under contract ED-IES-13-C-0017, Task 9, Option 9(C) at the request of the National Assessment Governing Board.

Introduction

Starting in early 2003, the National Assessment Governing Board embarked on an ambitious mission to redesign grade 12 assessments and reporting as recommended by the National Commission on 12th Grade Assessment and Reporting. Most importantly, the commission recommended that a state program should be implemented (similar to 4th and 8th grade) and that NAEP should start reporting on the readiness of 12th graders for college, training for employment, and entrance into the military. As a result of the second recommendation, a number of studies were conducted to assess whether and in what ways NAEP could report on *academic preparedness*. To be "academically prepared for college", 12th graders should have the knowledge and skills in reading and mathematics to qualify for placement into entry-level, credit-bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institution. After various content alignment studies, judgmental standard setting, secondary analyses, data collections, and statistical linking research (National Assessment Governing Board, 2009), potential benchmarks were identified on the 12th grade Reading and Mathematics scales to indicate what level of performance would correspond to a reasonable probability of being academically prepared for postsecondary education. As a result, a national postsecondary education preparedness percentage could be estimated and reported for the 2013 assessments in Reading and Mathematics. Details about this work can be found on a section of the National Assessment Governing Board website dedicated to preparedness (http://www.nagb.gov/what-we-do/commission.html).

As part of the initial statistical linking research, Florida participated (and continues to participate) at the 12th grade level and was a critical component for the validity evaluation of the benchmarks offering SAT®/ACT® data, Grade Point Averages, and ACCUPLACER® College Placement Exam results as well as longitudinal data into Florida public postsecondary institutions, including Remedial Course Placement and First Year Grade Point Average.

Moving forward, one focus of the second phase of the NAEP academic preparedness research is to study the extent to which grade 8 students are on track for being academically prepared for college by the end of high school. Several states, including Tennessee, participated in the statistical linking research and provided data on students who were part of the NAEP grade 8 sample during the 2012-2013 school year. Some state partners will continue to provide longitudinal data as these students progress through high school and beyond, to be analyzed and reported in future reports.

In this report we will describe the NAEP and EXPLORE® assessments in Reading and Mathematics, discuss the linking methodology (and refer the interested reader to more technical references), and provide the results. A summary will complete this report.

Linking Assessments

The ACT EXPLORE® Assessment

The EXPLORE® test[8] developed by ACT was administered to nearly all 8th graders in Tennessee during the 2012-2013 school year (with the testing window in Sep through Nov, 2012). The assessment includes four multiple-choice tests. Each test measures student's achievement in one of the following four areas: English, Mathematics, Reading, and Science. Students had 30 minutes to finish each test. The number of items in the test varies by subject, for reading and mathematics, both tests have 30 items. EXPLORE® scores provide evidence about the knowledge and skills that students are likely to have in each of the four aforementioned areas. The distribution of item difficulties was selected so that the tests will effectively differentiate among students who vary widely in the level of achievement. A composite score is provided, which is calculated as the average of the four test scores. The individual test scores, as well as the composite score, range from 1 to 25 and are disseminated to students and schools directly. In this study, only the Reading and Mathematics scores were used to link with the NAEP Reading and Mathematics assessments.

The ACT EXPLORE® assessments were designed to assess a specific student's academic progress at the 8th and 9th grade levels, especially with respect to college and career readiness. To help students translate test scores into a clear indicator of their current level of college readiness, ACT derived the ACT College Readiness Benchmarks based on a review of normative data, college admissions criteria, and information obtained through ACT's Course Placement Services. Students who meet a benchmark on the ACT test have approximately a 50% chance of obtaining a B or higher and approximately a 75% chance of obtaining a C or higher in the corresponding credit-bearing first-year college courses (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). In addition, there are corresponding benchmarks for the ACT EXPLORE®, which are linked to the ACT College Readiness Benchmarks. Students who meet a benchmark on the EXPLORE® test have approximately a 50% chance of meeting the ACT Benchmark in the same subject, and are likely to have approximately the same chance of earning a B or better grade in the corresponding college course(s) by the time they graduate high school. The current College Readiness Benchmarks for the EXPLORE® Reading test for grade 8 is 16 and for the EXPLORE® Mathematics test is 17 (ACT EXPLORE® 2013/2014 Technical Manual, p. 17). These benchmarks were used in this investigation. Note that the EXPLORE® reading benchmark was adjusted in 2013. Previously the reading EXPLORE® benchmark was 15. The math EXPLORE® benchmark remained unchanged.

[8] ACT will discontinue the use of the EXPLORE® test after fall 2015 for existing users and no new users are now being accepted.

The National Assessment of Educational Progress (NAEP)

The NAEP test was administered to selected 8th graders in Tennessee during the 2012-2013 school year (with the testing window from the last week of Jan to the first week of Mar in 2013). NAEP is the only nationally representative assessment of 4th, 8th, and 12th grade students in public and private schools in the U.S. in a variety of academic subjects. Subjects such as Reading, Mathematics, and Science are also assessed at the state- and even large urban district-level, particularly in grades 4 and 8. Samples of schools and students are selected from a sampling frame in order to produce results that are nationally representative and also representative of participating states and urban districts. Selected students had 50 minutes to complete the cognitive items (i.e., test questions) contained in the NAEP test booklets that were randomly assigned to them. The number and type of items in each booklet vary by subject and by grade. For grade 8 reading, each booklet contains two blocks of about 10 items each. For grade 8 math, each booklet contains two blocks of about 15 items each. A mix of multiple-choice and constructed response items is administered and blocks are systematically paired across booklets (i.e., matrix sampling design). The NAEP assessment is based on broad frameworks developed by the National Assessment Governing Board. By law, no student or school results are estimated or reported using the NAEP assessment. In fact, the assessment is designed in a way that no reliable score *can* be computed at the student level while minimizing the burden of any individual student selected to participate in the assessment. Instead, the main objective of NAEP is to report on the achievement of policy-relevant population groups, estimated directly using marginal estimation latent regression methods. For a comprehensive description of NAEP estimation procedures, the reader is referred to Mislevy, Beaton, Kaplan, & Sheehan (1992).

For the linking study, this requires that the relationship between NAEP and other measures (e.g., EXPLORE® scores) must be directly estimated using this latent regression methodology since there are no appropriate student-level scores available. In the methodology section we will discuss some of the steps that were required to complete this part of the research. NAEP reports results on scales that range from 0 to 500 in grade 8 Reading and Mathematics and the goal is to express the aforementioned ACT EXPLORE® benchmarks in terms of these scales. Students sampled for participation in NAEP are assessed in only one assessment subject. Consequently, each student in the matched or linking sample had EXPLORE® scores in both reading and mathematics, but results for only one NAEP assessment, either reading or mathematics.

Linking

When linking scales of different assessments, it is important to be precise about what that exactly entails. Usually, the two instruments under a linking study do not measure the same construct and have not been designed for that purpose, but generally there is some overlap. The greater the overlap, as evidenced by a higher correlation between the two scales, the more confident we can be that the instruments can be used to predict each other well. When the relationship is very strong and the instruments have a similarly high reliability, we would be able to claim that the two scales are

largely interchangeable and, therefore, that there is a one-to-one relationship between scores on the one scale and scores on the other scale. When this relationship is moderate, then we can do a 'best' projection of one scale onto the other or the reverse, which would not necessarily lead to similar results. In that case, the outcome would be of a probabilistic nature (e.g., "at score level X, students have a reasonably high probability to be prepared"). In the case of the preparedness linking studies, and taking past studies into account, a moderate relationship is most probable. We will elaborate further on this in subsequent sections.

Typically, a content alignment precedes statistical alignment to assess the extent to which the instruments were designed to measure the same or different constructs. Content alignment studies between NAEP and EXPLORE® Reading and Mathematics are being conducted by the National Opinion Research Center (NORC) at the University of Chicago (under contract ED-NAG-14-C-0002 with the National Assessment Governing Board) and will provide an important context for the statistical linking results presented here.

Methodology

In this section we will discuss the data and the linking methodology. The purpose is to give the reader some insight into the procedures that were followed and, therefore, the opportunity to evaluate the results within that context.

Data

This study used data from students who were sampled and assessed in NAEP 8th grade reading or mathematics in 2013 and had also taken the EXPLORE® assessment. From late January through early March of 2013, NAEP assessments in reading and mathematics were administered to samples of 8th grade students that were representative of each state, and together of the nation. As a result, about 2,700 public school students in Tennessee were sampled for each subject. Sample sizes are rounded to the nearest hundred as required in the NCES Statistical Standards (https://nces.ed.gov/statprog/2002/stdtoc.asp). Because only a sample is assessed and for efficiency purposes schools are sampled proportionally to size (in addition to other adjustments), sampling weights have to be used to appropriately represent all student groups of interest and, consequently, calculate unbiased results. The EXPLORE® assessment is required in Tennessee at the 8th grade level, meaning that almost all students who were sampled for NAEP also participated in EXPLORE® and have associated scores. The reverse is obviously not true, given that NAEP is sample-based (i.e., not every student who participated in EXPLORE® also participated in NAEP).

The process of matching EXPLORE® scores to NAEP participants was carried out through an agreement between the National Assessment Governing Board and the National Center for Education Statistics (NCES) to have NAEP contractors Westat and ETS conduct the preparedness

research work. In addition, data confidentiality agreements were established between all parties involved and the Tennessee Department of Education. A process for matching the student records was developed to protect students' identity and confidentiality. Confidentiality of state supplied scores (e.g., EXPLORE® scores) was assured through the assignment of a pseudo ID for students taking that assessment and using that pseudo ID as a way to transfer scores to ETS *without* the need to include Personally Identifiable Information (PII) such as names or birthdates. Similarly, the pseudo ID was appended to NAEP files by Westat who then provided that file to ETS, again *without* any PII. Via the pseudo ID, ETS subsequently matched EXPLORE® scores to NAEP files. In the case of Tennessee, EXPLORE® scores were matched at 93% for reading and 94% for mathematics, which is extraordinarily high. The matching rates for various student subgroups (by gender, by race/ethnicity, etc.) were at or above 91%. Table 1 provides weighted percentages by gender and race/ethnicity for the matched sample and overall match rates.

Table 1. Weighted percentages by gender and race of the Tennessee linking samples

Reading								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	36%	10%	3%	1%	#[1]	#	#	**51%**
Female	34%	11%	3%	1%	#	#	#	**49%**
***Total*[2]**	**71%**	**21%**	**6%**	**1%**	**#**	**#**	**1%**	**100%**
							Overall Match Rate	**93%**
Mathematics								
	White	*Black*	*Hispanic*	*Asian*	*American Indian /Alaskan Native*	*Pacific Islander*	*2+ races*	***Total*[2]**
Male	36%	11%	3%	1%	#	#	#	**51%**
Female	35%	10%	3%	1%	#	#	#	**49%**
***Total*[2]**	**71%**	**21%**	**6%**	**2%**	**#**	**#**	**#**	**100%**
							Overall Match Rate	**94%**

NOTES: [1]# Rounds to zero.
[2] Detail may not sum to totals because of rounding.

Given the fact that the two assessments that are linked have very different purposes and, possibly, different stakes, an outlier analysis is in order. For instance, if there are participants that scored very high on a *higher* stakes test (i.e., EXPLORE® test) and very low on the *lower* stakes test, the low performance can be reasonably attributed to motivation rather than performance level. Such cases would be considered 'outlier' and removed from further analyses. An initial examination of the joint distribution of NAEP and EXPLORE® revealed very few potential outlier cases. After this more

cursory inspection, standardized residuals from robust regression (Huber, 1973) were used to identify approximately 0.6% of cases in reading and approximately 0.8% of cases in mathematics (cases with absolute standardized residuals greater than 3 were considered outliers and removed). We refer to Huber (1973) for details about the procedure and the criteria applied. These outliers were excluded from the final linking samples and were not used in subsequent analyses.

Analysis Approach

After preparatory data identification, matching, merging, and data reconciliation, the linking analyses were conducted. The current study was designed to pursue three specific analysis questions that guide the choices in methodology for the linking and validation:

1) What are the correlations between the grade 8 NAEP and EXPLORE® scores in reading and mathematics?
2) What scores on the grade 8 NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks?
3) What are the average grade 8 NAEP reading and mathematics scores (and the difference between the 75th and 25th percentiles) and the IQR for students below, at, and above the EXPLORE® benchmarks?

We will describe pertinent methodological details about the analysis followed by the results of the analyses in the final section. The key steps of the analysis are (a) estimating the correlation between NAEP and EXPLORE®, which includes use of the aforementioned latent regression methodology (b) determining the appropriate methodology for linking based on those correlations (c) applying procedures to effectively estimate cumulative probability functions and (d) calculating impact data as part of the results.

A satisfactory treatment of the latent regression methodology is outside the scope of this report and the interested reader is referred to Mislevy, Beaton, Kaplan, and Sheehan (1992). The basic notion is that NAEP measures constructs that are represented on item response theory based latent scales, which are not measured reliably at the student level. However, pertinent data from students in specified groups of interest can be pooled to estimate reliable scores at the group level. EXPLORE® scores, on the other hand, are reliably estimated at the individual level and can be treated as a set of consecutive (semi-continuous) groups. Correlations between NAEP and EXPLORE® can be directly estimated at the overall level and the result showed that the (true score) correlation for reading is 0.73 and for mathematics is 0.81. While these are not low correlations, they do suggest that there is enough uncertainty in the relationship that a direct one-to-one correspondence of scale score points is not advisable.

To elaborate on that observation and as briefly introduced earlier, different classes of statistical relationships can be established between various tests, and the distinctions correspond to the extent

to which the tests are similar with respect to the constructs measured, populations, and measurement characteristics of the tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Holland & Dorans, 2006). In this study, two types of statistical linking were originally considered: concordance and projection. Concordance establishes a score linkage between two tests by matching the corresponding score distributions. The claims that can be made based on concordance are also commensurately strong. Essentially, the claim is made that a score *x* on NAEP exactly corresponds to a score *y* on EXPLORE® and vice versa. Projection is a less stringent type of correspondence in which scores on one test are related, typically via a linear or nonlinear regression, to a conditional distribution of scores on the other test. Projection relationships are not symmetric, and do not assume or result in a one-to-one correspondence. The claim is made that a score of *x* on NAEP corresponds to the proportion *p* of students attaining the benchmark score of *y* or higher on EXPLORE®. Subsequently, a choice for *p* has to be made, where a more conservative claim requires a higher *p*. This means that if one wants to have a very high degree of confidence that students at a certain NAEP score pass the benchmark, then a relatively high *p* has to be set, a relatively high score level is identified, and, likely, the percent of students that actually pass the benchmark is under-estimated. The reverse is true when a lower degree of confidence is acceptable. Needless to say, concordance assumes and requires a much stronger relationship than projection.

The relationships between NAEP and EXPLORE® reading (r =0.73) and mathematics (r = 0.81) are not sufficiently strong to support concordance, given that a generally accepted minimum correlation for concordance is r = 0.866 (Dorans, 1999; Dorans & Walker, 2007). Consequently, projection was used in this study. As mentioned before, typically a smoothing process is applied in order to produce more accurate probability distributions, particularly when the underlying population distribution of test scores may contain irregularities (Moses & Liu, 2011), for example due to a non-continuous nature of the scale. Bivariate loglinear smoothing (Holland & Thayer, 2000) was applied to the joint NAEP-EXPLORE® distributions[9].

An important tool for evaluating statistical links between tests is sensitivity analysis, which is intended to examine the extent to which the linking relationship is invariant across key student groups, such as gender and race/ethnicity groups. These analyses require a minimum sample size[10] in order to produce reliable comparisons. For the Tennessee linking samples, both gender groups met that criterion. For the race/ethnicity groups, only White and Black student subgroups met the criterion. Separate linking functions were established for these subgroups and deviations from the

[9] For reading, as part of the loglinear smoothing procedure we preserved the first 3 moments for the NAEP distribution, 4 moments for the EXPLORE® distribution, and 4 cross-moments. For math, we preserved the first 6 moments for the NAEP distribution, 6 moments for the EXPLORE® distribution, and 4 cross-moments. These loglinear smoothing models mostly resulted in the smallest value of the Akaike Information Criterion (AIC) statistic (Moses & von Davier, 2006), although model complexity and sample size was also taken into consideration.

[10] The minimum was set at 500 as a rule of thumb, but based on the idea that there is at least one observation below -3 and above +3 standard deviations (in a standard normal distribution) in expectation.

overall linking function indicated violation of invariance. It should be noted though that the purpose of this linking is to establish a specific benchmark for preparedness. In that sense, substantial variability across student groups for parts of the scale that does not entail the benchmark could be quite harmless. For NAEP reading, no substantial deviation from the overall linking function was detected for Male, Female, or White student subgroups. The linking function for Black students was slightly lower than the overall linking function. For NAEP math, no substantial deviation from the overall linking function was detected for Female or White student subgroups. The linking functions for Male and Black students were slightly lower than the overall linking function. Even though the comparison between the linking functions indicated some variance among different subgroups, the difference was not large enough to discredit the linking study. In fact, it should be emphasized that some subgroups considered here had a much smaller sample size than the overall linking sample, and therefore the difference observed between the linking functions should be interpreted with great caution.

Finally, for both reading and mathematics, the probabilities from the smoothed joint distributions were used to create projections tables containing conditional cumulative distributions of NAEP proficiencies for EXPLORE® scores. The range of possible NAEP scores below, at, and above the EXPLORE® benchmark (16 on the EXPLORE® reading scale and 17 on the EXPLORE® mathematics scale) were estimated and, subsequently, for each subject area the projected conditional distributions were used to identify the NAEP scale scores associated with the EXPLORE® benchmarks.

In the following section we will discuss the results of the linking study, focusing on the second and third analysis questions: What NAEP scores correspond to the EXPLORE® benchmarks and what are the distributional characteristics associated with those benchmarks.

Results

On Track Markers

The most important result, following the second and third analysis questions, is to determine what scores on the NAEP reading and mathematics scales correspond to the EXPLORE® benchmarks. In other words, what would be the 'on track to be prepared' score level on NAEP that corresponds most reasonably to an established 'on track' benchmark.

Table 2 provides descriptive statistics to get an initial sense of where the benchmark most likely will be located as well as some distributional properties as context to these results. The average scores and percentile estimates for students below, at, and above the EXPLORE® benchmarks are spread out, though more so for students below the benchmark than above. Note that the mean *at* the benchmark is not necessarily the same as the NAEP score equivalent for the benchmark, but rather a

characterization of the students at this level. Also note that these results are based on the statistical linking (i.e., projection methodology).

Table 2: Descriptive NAEP Statistics for Students Below, At, or Above the EXPLORE® Benchmarks

Subject	EXPLORE® Benchmark	Mean	Percentage[2]	SD	Percentile		IQR[1]
					25th	*75th*	
Reading	*Below*	252	64%	28	234	271	37
	At	279	8%	22	264	293	29
	Above	297	27%	21	282	311	29
Mathematics	*Below*	264	65%	28	247	284	37
	At	294	12%	18	282	306	24
	Above	317	23%	22	301	331	30

NOTES: [1]IQR is the Inter Quartile Range or the difference between the 75th and 25th percentiles.
[2] Detail may not sum to totals because of rounding.

To determine the NAEP scale score point that most reasonably corresponds to the EXPLORE® benchmarks, it is most illustrative to graphically represent the relationship. Figures 1 and 2 show the relationship based on statistical projection for students at the respective benchmarks. The black curved line shows the proportion of students meeting the EXPLORE® benchmark for pertinent score levels on NAEP. Colored vertical lines indicate where the NAEP achievement levels are located. Finally, and as mentioned before, a proportion level has to be chosen commensurate the confidence required to indicate whether students have passed the benchmark or not. A red dotted line shows at which point students are more likely to have reached the benchmark than not (i.e., the probability is set at 0.50). Given the moderate relationships between the two scales, this seems a reasonable location for indicating sufficient chance to be 'on track to preparedness'. For context, a secondary, lighter red line indicates when the probability is set at 0.80, indicating a relatively high level of confidence that students have attained the EXPLORE® benchmark.

From the graphs it can be deducted that the location where students have a reasonable probability to be on track for reading could be set at a NAEP scale score of 284, slightly above the *Proficient* achievement level. The mathematics counterpart could be set at 296, slightly below the *Proficient* achievement level.

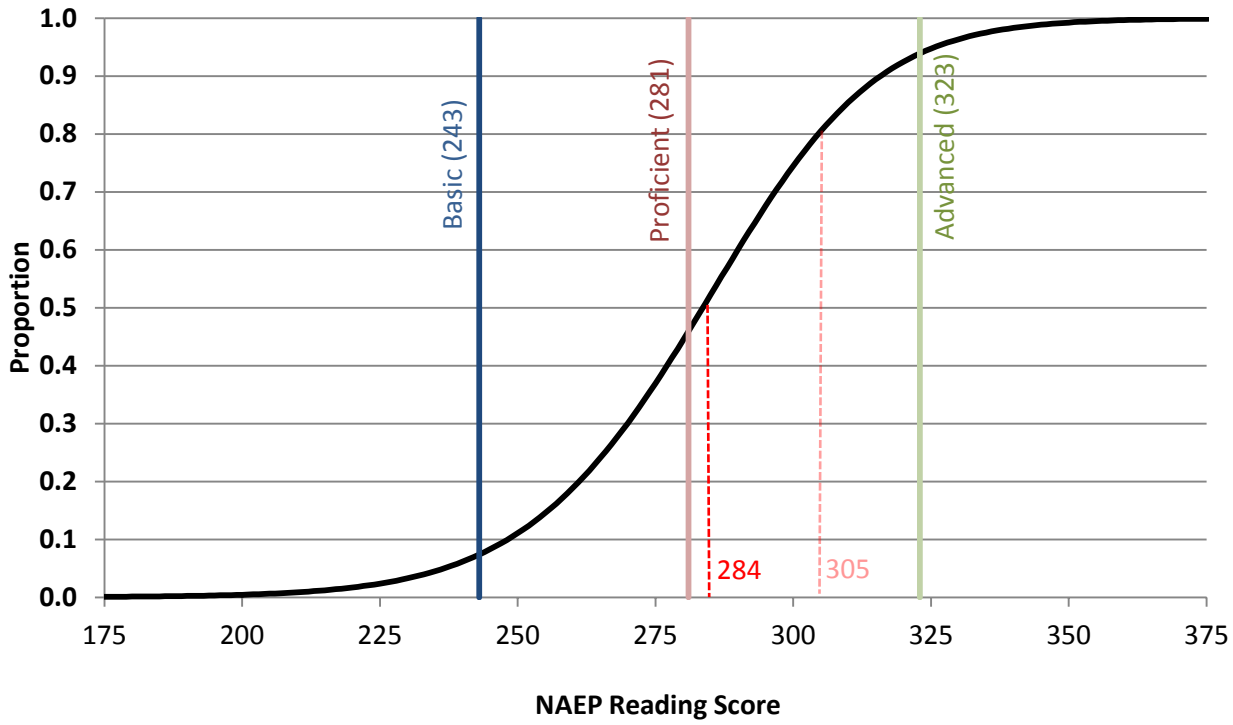

Figure 1: Proportion of students meeting the Reading EXPLORE® benchmark of 16 in Tennessee for NAEP Reading levels

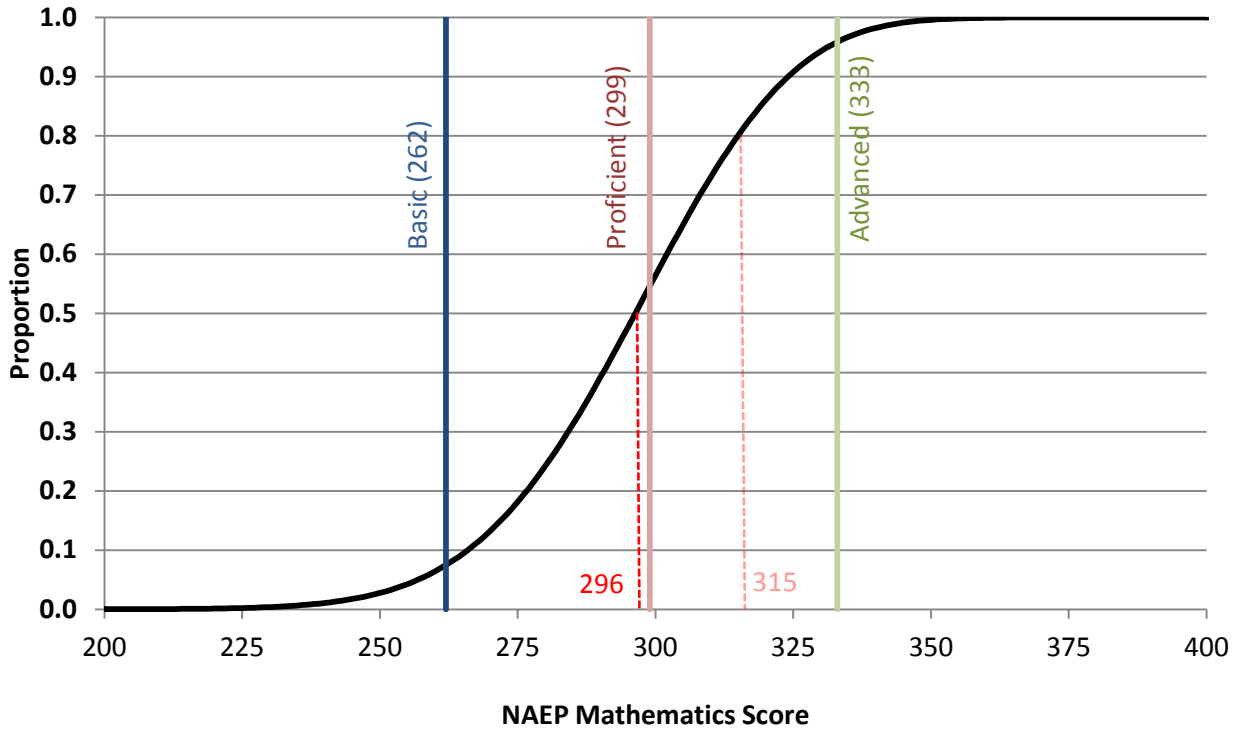

Figure 2: Proportion of students meeting the Mathematics EXPLORE® benchmark of 17 in Tennessee for NAEP Mathematics levels

Impact

Now that potential points have been identified, it is important to show what percentage of students in Tennessee are deemed to have a reasonable probability (i.e., the probability set at 0.50) of being on track in grade 8 across various student groups. Table 3 provides those percentages, based on the potential points identified on the NAEP scales, as well as the EXPLORE® benchmarks. Table 3 indicates that overall about 31 to 36 percent of students are on track, but the results differ across different subgroups. No significance testing has been conducted to compare these percentages and, therefore, no comparative statements will be made.

Table 3: Percentage of the Tennessee linking samples that have a reasonable probability to be on track to be academically prepared based on the potential points identified on the NAEP scale, compared to the percentage of the same sample meeting the Reading EXPLORE® benchmark of 16 and Mathematics EXPLORE® benchmark of 17.

	Reading		Mathematics	
Student Group	**NAEP ≥ 284**	**EXPLORE® ≥ 16**	**NAEP ≥ 296**	**EXPLORE® ≥ 17**
Total	31%	33%	32%	36%
Male	27%	31%	33%	35%
Female	34%	35%	31%	36%
White	35%	38%	37%	41%
Black	15%	17%	12%	18%
Hispanic	26%	26%	26%	29%

Summary

The goal of this study was to statistically relate NAEP and EXPLORE® and use that relationship to identify a reference point or range on the NAEP 8th grade reading and mathematics scales reasonably associated with ACT's preparedness benchmarks on the EXPLORE® reading and mathematics measures. Identifying such points would potentially allow NAEP to report on the percentage of students at 8th grade who are on track to be prepared for college for the nation and for states. The first step involves three participating states, including Tennessee, who have graciously provided the critical EXPLORE® data necessary to calculate the relationship with NAEP. In this study, various statistical techniques, including latent regression, smoothing, and statistical projection were used to establish the relationship and identify potential markers on the NAEP scale that could form the basis for 'on track to preparedness' reporting (see Figures 1 and 2 for examples of how the markers were determined).

A key finding was that the relationship between the two scales is moderate, meaning that the kind of relational statements that can be made need to be presented in notions of probability rather than direct one-to-one relationships. This is not surprising because the instruments are not intended to measure the exact same construct, however, it does make interpretation somewhat more challenging. The results showed that NAEP scale score points just above the *Proficient* achievement levels could form a reasonable basis for reporting 'on track for preparedness'. Approximately 31% of Tennessee 8th graders met that criterion for reading and 32% met the criterion for math. Further content alignment work, which is conducted independently from this study, should provide further context to these results.

References

ACT EXPLORE Technical Manual 2013/2014. (http://www.act.org/explore/pdf/TechManual.pdf)

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Research Report No. 99-2). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp. 179-198). New York: Springer.

Feuer, M.J., Holland, P.W., Green, G.F., Bertenthal, M.W., & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington, DC: National Academy Press.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Washington, DC: American Council on Education.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29* (2), 133-161.

Moses, T.P., & Liu, J. (2011). *Smoothing and Equating Methods Applied to Different Types of Test Score Distributions and Evaluated With Respect to Multiple Equating Criteria* (Research Report No. 11-20). Princeton, NJ: Educational Testing Service.

Moses, T. P., & von Davier, A. A. (2006). *An SAS macro for loglinear smoothing: Applications and implications* (Research Report No. 06-05). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*.

Evaluation of NAEP Achievement Levels

Objective To receive a brief informational update on the current status of the independent evaluation of NAEP achievement levels that is being performed by the National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES). Ongoing updates will be provided at each COSDAM meeting.

Background

The NAEP legislation states:

> The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public.

In providing further detail, the aforementioned subsection (f) outlines:

(1) REVIEW-

A. IN GENERAL- The Secretary shall provide for continuing review of any assessment authorized under this section, and student achievement levels, by one or more professional assessment evaluation organizations.

B. ISSUES ADDRESSED- Such continuing review shall address--

(i) whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);

(ii) whether student achievement levels are reasonable, valid, reliable, and informative to the public;-

(iii) whether any authorized assessment is being administered as a random sample and is reporting the trends in academic achievement in a valid and reliable manner in the subject areas being assessed;

(iv) whether any of the test questions are biased, as described in section 302(e)(4); and

(v) whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

(2) REPORT- The Secretary shall report to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health, Education, Labor, and Pensions of the Senate, the President, and the Nation on the findings and recommendations of such reviews.

(3) USE OF FINDINGS AND RECOMMENDATIONS- The Commissioner for Education Statistics and the National Assessment Governing Board shall consider the findings and recommendations of such reviews in designing the competition to select the organization, or organizations, through which the Commissioner for Education Statistics carries out the National Assessment.

Evaluation of NAEP Achievement Levels Contract

The National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES), will administer the Evaluation of the NAEP Achievement Levels. On September 29, 2014, NCEE awarded a contract to The National Academy of Sciences to perform this work.

Objectives for the evaluation include the following:

- Determine how "reasonable, valid, reliable and informative to the public" will be operationalized in this study.
- Identify the kinds of objective data and research findings that will be examined.
- Review and analyze extant information related to the study's purpose.
- Gather other objective information from relevant experts and stakeholders, without creating burden for the public through new, large-scale data collection.
- Organize, summarize, and present the findings from the evaluation in a written report, including a summary that is accessible for nontechnical audiences, discussing the strengths/ weaknesses and gaps in knowledge in relation to the evaluation criteria.
- Provide, prior to release of the study report, for an independent external review of that report for comprehensiveness, objectivity, and freedom from bias.
- If the optional tasks are authorized by ED, plan and conduct dissemination events to communicate the conclusions of the final report to different audiences of stakeholders.

Design:

This study will focus on the achievement levels used in reporting NAEP results for the reading and mathematics assessments in grades 4, 8, and 12. Specifically, the study will review developments over the past decade in the ways achievement levels for NAEP are set and used and will evaluate whether the resulting achievement levels are "reasonable, valid, reliable, and informative to the public." The study will rely on an independent committee of experts with a broad range of expertise related to assessment, statistics, social science, and education policy. The project will receive oversight from the Board on Testing and Assessment (BOTA) and the Committee on National Statistics (CNSTAT) of the National Research Council.

Members of the interdisciplinary review committee were selected in early 2015 (see below), and the committee is expected to meet over the course of 2015. The report from the evaluation is expected to be released in 2016 and will be announced on http://ies.ed.gov/ncee/.

Name	**Affiliation**
Dr. Christopher F. Edley, Jr. (Chair)	University of California, Berkeley
Dr. Peter Afflerbach	University of Maryland, College Park
Dr. Sybilla Beckmann	University of Georgia
Dr. H. Russell Bernard	University of Florida
Dr. Karla Egan	National Center for the Improvement of Educational Assessment
Dr. David J. Francis	University of Houston
Dr. Margaret E. Goertz	University of Pennsylvania
Dr. Laura Hamilton	The RAND Corporation
Dr. Brian W. Junker	Carnegie Mellon University
Dr. Suzanne Lane	University of Pittsburgh
Ms. Sharon J. Lewis	Retired
Dr. Bernard L. Madison	University of Arkansas
Dr. Scott Norton	Council of Chief State School Officers
Dr. Sharon Vaughn	The University of Texas at Austin
Dr. Lauress L. Wise	HumRRO

Additional information about the Committee and project activities is available at: http://www8.nationalacademies.org/cp/projectview.aspx?key=49677. The first Committee meeting took place in Washington, DC on February 19-20, 2015. Governing Board staff attended the open session and made a presentation to the Committee on the history of the NAEP achievement levels setting activities. The second meeting of the Committee took place in Washington, DC on May 27-28, 2015. Governing Board staff attended the open session on the afternoon of May 27th to listen to panel discussions about interpretations and uses of NAEP achievement levels (see attached agenda).

INTERPRETATIONS AND USES OF NAEP ACHIEVEMENT LEVELS

May 27, 2015,
1:00-5:00

Committee on the Evaluation of NAEP Achievement Levels in Reading and Math

National Academy of Science Building
Lecture Room
2101 Constitution Ave., NW
Washington DC

AGENDA

This session is sponsored by the National Academy of Science's Committee on the Evaluation of NAEP Achievement Levels in Reading and Math, which is charged with evaluating the extent to which NAEP achievement levels are reasonable, reliable, valid, and informative to the public. The Committee's goal for the session is to gather information on uses/interpretations of NAEP results that will help to guide their evaluation.

The session is separated into 5 parts, each led by a group of panelists from a variety of perspectives. The panel discussions will each be facilitated by a committee member, with the goal of having a free-flowing, moderated conversation among the panelists, audience, and committee members.

1:00 **WELCOME, OVERVIEW OF AGENDA**
Brian Junker, Carnegie Mellon, Committee Member

1:10 **PANEL DISCUSSION 1: EDUCATION WRITER PERSPECTIVES**
Facilitator: Brian Junker, Carnegie Mellon, Committee Member

- **Sarah Butrymowicz**, Hechinger Report
- **Catherine Gewertz**, Education Week
- **Lyndsey Layton,** Washington Post
- **Emily Richmond**, Education Writers Association
- **Bob Rothman,** Alliance for Excellent Education

1:55 **PANEL DISCUSSION 2: STATE AND LOCAL POLICY PERSPECTIVES**
Facilitator: Scott Norton, CCSSO, Committee Member

- **Michael Casserly,** Council of Great City Schools
- **Scott Jenkins,** National Governors Association
- **Wendy Geiger**, Virginia Department of Education
- **Nathan Olson**, Washington Department of Education

2:40 **Break**

2:55 **PANEL DISCUSSION 3: EDUCATION POLICY AND ADVOCACY PERSPECTIVES**
Facilitator: Laura Hamilton, RAND, Committee Member

- **Patte Barth**, National School Board Association
- **Renee Jackson**, National PTA
- **Sonja Brookins Santelises**, Education Trust
- **Dara Zeehandelaar,** Fordham Institute

3:40 **PANEL DISCUSSION 4: USES OF NAEP ACHIEVEMENT LEVELS FOR ASSESSMENTS OF THE COMMON CORE STATE STANDARDS**
Facilitator: Suzanne Lane, University of Pittsburgh, Committee Member

- **Enis Dogan,** Partnership for Assessment of Readiness for College and Careers (PARCC)
- **Jacqueline King,** Smarter Balanced Assessment Consortium

4:10 **PANEL DISCUSSION 5: SYNTHESIS**
Facilitator: Brian Junker, Carnegie Mellon, Committee Member

- **Michael Kane**, ETS
- **Lorrie Shepard**, University of Colorado

4:50 **Wrap Up, Final Q&A**

5:00 **Adjourn open session**

NAEP Job Training Preparedness Report

During the past 10 years, the Governing Board has commissioned more than 30 research studies to investigate whether the 12th grade NAEP reading and mathematics assessments could serve as indicators of students' academic preparedness for college and job training. The research results supported the claim that 12th grade NAEP assessments of reading and mathematics are indicators of academic preparedness for college. However, in the area of job training, the research studies have not supported the use of NAEP as an indicator of job training preparedness.

Given the prominence of career-readiness discussions across the country, it was determined that a synopsis of the Board's extensive job training preparedness research would be of interest to the field.

The purpose of this report is to summarize the context, methodology, results, and conclusions of the Governing Board's job training preparedness research studies for NAEP. The types of job training research studies include content alignment, judgmental standard setting, and other areas. This report is being written for educators, policymakers, researchers, and interested members of the general public. Therefore, this report is not intended to provide the full details of each study as those are fully documented on the Board's 12th Grade Preparedness Technical Report website (http://www.nagb.org/what-we-do/preparedness-research.html). For those who wish to review the studies and results in detail, links to the individual research study reports will be embedded in the body of the job training preparedness summary report.

The Job Training Preparedness Report is being developed by Widmeyer Communications, under Governing Board contract ED-NAG-11-O-0005 for preparedness reporting. A draft of the report is currently being reviewed by Governing Board staff. In the fall a subsequent draft will be sent to COSDAM for review, and the final report will appear in the November 2015 COSDAM materials.

Procurement Update

Review of Existing Studies on Motivation and Engagement in NAEP

During the August 2013 COSDAM meeting, former Governing Board Executive Director Cornelia Orr reported on the desk side briefings that she had given to policy leaders and organizations about the results of the Governing Board's academic preparedness research. Ms. Orr reported that one of the questions she received was about whether grade 12 students are motivated to try hard on NAEP. Ms. Orr noted that it is important to be aware of the tendency to question whether grade 12 results represent students' best efforts. Some people have a hard time believing that 12th-graders try hard on a test that does not count. On the other hand, TIMSS and PISA are at the secondary level and also do not count.

There is some evidence that grade 12 students do take the test seriously, such as completion rates and completion of open-ended questions in particular. During the March 2014 COSDAM meeting, Samantha Burg of the National Center for Education Statistics (NCES) presented some encouraging data on grade 12 school and student participation rates and item response rates (from 1992 to 2013) and comparisons to grades 4 and 8. A Focus on NAEP report, *Grade 12 Participation and Engagement in NAEP*, is scheduled to be released by NCES in August 2015.

On the other hand, if an ERIC search was performed on the terms "NAEP" and "motivation," the search would yield studies that conclude students are not very motivated. Previous COSDAM discussions have noted that the secondary research on NAEP and motivation which is often cited has not been critiqued for technical merit. One idea discussed during previous COSDAM meetings is that a literature review and critique of existing studies could be performed as part of the efforts on preparedness research.

To pursue this idea, at the November 2014 COSDAM meeting, Committee members discussed a procurement to conduct a review and summary of existing research on motivation and engagement in NAEP, with the following goals:

- To critically evaluate the claims that have been made;
- To summarize the extent to which results are consistent across studies; and
- To recommend future research that could be performed.

A request for proposals was issued on June 18, 2015 and can be found at: https://www.fbo.gov/index?s=opportunity&mode=form&id=ff81f36e3d1393a1b83b79a56f0eb12f&tab=core&_cview=0. Proposals are due on July 21, 2015, with a target award date of September 2015.

The Performance Work Statement (PWS) includes a requirement for: a design document that lays out the proposed plan of research, as well as the process to identify studies to include in a critical synthesis of the extant research; identification of all extant studies, reports, papers, and

research relevant to the topic; an annotated summary of each of these studies; a critical evaluation of methods, claims, findings, conclusions among included studies; a comprehensive synthesis of findings from all included studies, especially the extent to which results may be common across studies; and recommendations for future research.

At the November 2015 meeting, COSDAM will receive an update about the status of the contract award, including project milestones completed.