Numeracy
Advancing Education in Quantitative Literacy

Volume 10 | Issue 2 Article 3

2017

Preliminary Evaluation of the Psychometric Quality of HEIghten™ Quantitative Literacy

Katrina C. Roohr
Educational Testing Service, kroohr@ets.org
HyeSun Lee
California State University Channel Islands, hyesun.kj.lee@gmail.com
Jun Xu
Educational Testing Service, jxu@ets.org
Ou Lydia Liu
Educational Testing Service, lliu@ets.org
Zhen Wang
Educational Testing Service, jwang@ets.org

Recommended Citation

Roohr, Katrina C., HyeSun Lee, Jun Xu, Ou Lydia Liu, and Zhen Wang. "Preliminary Evaluation of the Psychometric Quality of HEIghten™ Quantitative Literacy." *Numeracy* 10, Iss. 2 (2017): Article 3. DOI: http://doi.org/10.5038/1936-4660.10.2.3

Preliminary Evaluation of the Psychometric Quality of HEIghten™ Quantitative Literacy

Abstract

Quantitative literacy has been identified as an important student learning outcome (SLO) by both the higher education and workforce communities. This paper aims to provide preliminary evidence of the psychometric quality of the pilot forms for *HEIghten* quantitative literacy, a next-generation SLO assessment for students in higher education. We evaluated the psychometric quality of the test items (e.g., item analyses), individual- and group-level reliability, the relationship with student performance and related variables (e.g., grade point average) as well as student perceptions, and differences across college-related and demographic subgroups. Our study used data from a pilot test administered to over 1,500 students at 23 higher education institutions in the United States. Results showed that (a) overall, items were functioning well, but a small portion of items should be dropped due to unsatisfactory performance; (b) correlations across sub-areas of the assessment were very high indicating that the assessment may be unidimensional; (c) reliability estimates similar to existing SLO assessments were found at both individual and group levels; (d) assessment scores correlated positively with high school and college GPA, number of math college courses, self-rated quantitative literacy skills, and college admissions scores; (e) students had positive perceptions about the assessment; and (f) performance differences were found across institution type, college majors, gender, racial/ethnic groups, and language groups, but not across credit-hour categories. Implications for operational test development and understanding of quantitative literacy performance are discussed.

Keywords

quantitative literacy, assessment, higher education, student learning outcomes, reliability, validity

Cover Page Footnote

Katrina Crotts Roohr is a Managing Research Scientist at the Educational Testing Service. Her current work involves research on student learning outcomes in higher education. She has conducted research on assessing quantitative literacy, civic competency and engagement, oral communication, and critical thinking in higher education as well as research around student learning gains in college, and learning outcomes at community colleges. Katrina received her Ed.D. in Psychometric Methods, Educational Statistics, and Research Methods from the University of Massachusetts Amherst.

HyeSun Lee is currently a lecturer for the Psychology Program at California State University Channel Islands and will be joining the university as an assistant professor beginning August 2017. Her research lies in differential item functioning, multilevel modeling, propensity scores, and faking in noncognitive assessments. She received her Ph.D. at the University of Nebraska-Lincoln, specializing in Quantitative, Qualitative, and Psychometric Methods.

Jun Xu is a Research Data Analyst at the Educational Testing Service. She graduated from the University of Pennsylvania in 2012 with a master's degree in Educational Statistics. She is currently pursuing her Ph.D. in Statistics and Measurement at Rutgers University. Her current work concentrates on the use of a wide variety of data analysis methods, such as structural equation modeling, multivariate analyses, and complex sampling design to help foster research in many areas of education.

Ou Lydia Liu is a Senior Research Director at the Educational Testing Service. Her main research areas include assessment of student learning outcomes in higher education and innovative science assessment in K-12. Lydia has published widely in applied measurement, higher education, and science education journals, including Educational Researcher and Science. She was the recipient of the NCME Jason Millman Promising Measurement Scholar Award in 2011. Lydia holds a bachelor's degree in Science & English from the University of Science and Technology of China and a doctoral degree in Quantitative Methods and Evaluation from the University of California, Berkeley.

Zhen Wang is a Senior Psychometrician at the Educational Testing Service (ETS). She holds a Ph.D. in Educational Measurement and Research Methodology from the University of British Columbia. Currently, she is the statistical coordinator for ETS' HEIghten® Outcomes Assessment Suite, Major Field Test, Success Navigator, Proficiency Profile, and iSkills projects. She also works on the research projects related to automated scoring and higher education. Her research interests include item response theory, structural equation models, hierarchical linear models, rater models, and equating and scaling techniques.

Introduction

Quantitative literacy is the application and practical use of mathematics to real-world contexts (e.g., Sons 1996; Steen 2001; Rhodes 2010; Roohr et al. 2014). Also referred to as quantitative reasoning, quantitative fluency, mathematical literacy, or numeracy (Elrod 2015), quantitative literacy is distinct from traditional mathematics and goes beyond knowledge of formulas and equations (Steen 2001). Quantitative literacy is a "habit of the mind" enhancing mathematics, often focusing on the "logic of certainty" and involving data from the empirical world (Steen 2001, 5). Additionally, it involves the intersection of critical thinking and basic mathematics skills across various disciplines or real-world contexts (Elrod 2015).

Quantitative literacy is considered an important element to today's democratic society (Steen 2001; Shavelson 2008), and it has been recognized as an important student learning outcome (SLO) across the higher education and workforce communities. For instance, 91% of the chief academic officers at 433 colleges and universities across the United States identified quantitative reasoning as an important intellectual and practical skill (Association for American Colleges and Universities [AAC&U] 2011). The importance of quantitative literacy has also been echoed by the workforce community. In a recent survey by Hart Research Associates (2015), 56% of the 400 surveyed employers rated the ability to work with numbers and understand statistics as a very important SLO for college graduates entering the workforce.

Despite its importance, direct evidence from the Programme for the International Assessment for Adult Competencies (PIAAC) has shown that adults in the United States are underprepared to use quantitative skills. PIAAC results showed that only 18% of U.S. adults ages 16 to 65 with bachelor's degrees scored in the top two proficiency levels (out of five) on the Numeracy measure, compared to an international average of 24% (Goodman et al. 2013). Additionally, only 28% of the 400 surveyed employers by Hart Research Associates (2015) indicated that college graduates were well prepared to work with numbers and statistics, suggesting substantial room for improvement. With the increased importance of SLOs such as quantitative literacy, there is a critical need to evaluate whether students are developing these skills successfully prior to graduating college, regardless of college major (Dumford and Rocconi 2015).

Given the importance of quantitative literacy, it has been increasingly included as a key learning outcome by higher education institutions. One way to measure student learning in quantitative literacy is through the use of SLO assessments in higher education. An SLO assessment for quantitative literacy could provide an institution with information to identify gaps in students'

quantitative literacy performance and evaluate group-level performance at one or multiple time points, thus providing information that could help identify potential changes in the curriculum and instruction that may need to be made to ensure that students are prepared to use quantitative literacy skills upon graduating college. To ensure that an SLO assessment in quantitative literacy can be used for these various purposes, we first need to collect evidence to support these intended uses of student scores.

In this paper, we discuss the assessment design process and conduct a preliminary evaluation of the psychometric quality of a next-generation[1] SLO assessment, *HEIghten*™ (the capitalized HEI stands for Higher Education Institution). *HEIghten* Quantitative Literacy is a college-level assessment that evaluates students' abilities to comprehend, detect, and solve mathematics problems in authentic contexts (including personal and everyday life, workplace, and societal contexts) across a variety of mathematical content areas. The assessment is one module out of a five-assessment *HEIghten* Outcomes Assessment Suite[2] intended to measure general education SLOs for all college students, regardless of college major (the other assessment modules are critical thinking; written communication; intercultural competency and diversity; and civic competency and engagement). These generic-skills assessments are intended to be used mainly at the institution level, providing group-level information about student learning to inform regional and program accreditation, external accountability, curriculum modification, institutional improvement, and benchmark performance both externally and internally. That said, these assessments can also be used at the individual level to provide information about students' overall performance and performance levels in these various competencies.

Developing *HEIghten* Quantitative Literacy: Assessment Design

There are numerous stages in the test development process before an assessment can be implemented operationally. Some of the major steps are: (a) identifying the purpose of the assessment; (b) developing and evaluating the test specifications; (c) developing, testing, evaluating, and modifying the test items; (d) assembling

[1] When using the term "next-generation assessment," we are referring to an assessment that is: (a) administered online using technology-enhanced items that go beyond traditional single-selection multiple-choice items, (b) developed based on a theory-driven framework that's aligned with up-to-date research, (c) of high psychometric quality, and (d) based in real-life contexts.

[2] https://www.ets.org/heighten "Introducing the *HEIghten* Outcomes Assessment Suite" (accessed May 23, 2017)

DOI: http://doi.org/10.5038/1936-4660.10.2.3

the test forms; and (e) developing the procedures and materials for administration and scoring (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 2014). In all stages of test development, it is important also to consider validity, reliability, and fairness (AERA et al. 2014). In this study we focus on the third step of test development: developing, testing, evaluating, and modifying test items to ultimately inform the assembly of operational test forms. However, it is first important to discuss the first two steps that were conducted to develop *HEIghten* Quantitative Literacy, a new assessment designed to reflect the latest advancements in research and assessment of college-level quantitative knowledge and skills.

An Evidence-Centered Design Approach

To develop *HEIghten* Quantitative Literacy, we took an evidence-centered design (ECD) approach (Mislevy et al. 2003). ECD provides a structural framework for developing assessments. In this framework, we first determined what construct (i.e., knowledge, skills, or attributes) and dimensions (or aspects) of that construct should be assessed (i.e., the student model), what behaviors or performances should reveal those constructs (i.e., evidence models), and what tasks should elicit those behaviors (i.e., task models). Using this information, we then evaluated how these three models work together to form an assessment (i.e., the assembly model; Mislevy et al. 2002).

The ECD approach for developing *HEIghten* Quantitative Literacy is discussed in detail in Roohr et al. (2014). To determine what knowledge, skills, or attributes should be assessed, Roohr and colleagues (2014) reviewed existing definitions, frameworks, and assessments of quantitative-related constructs (e.g., quantitative literacy, quantitative reasoning, numeracy, mathematical literacy, etc.) in higher education and the workplace. Some of the existing frameworks and definitions included AAC&U's VALUE (Valid Assessment of Learning in Undergraduate Education) rubrics, Lumina's Degree Qualifications Profile (DQP), the Mathematical Association of America, Organisation for Economic Co-Operation and Development (OECD), and the American Mathematical Association of Two-Year Colleges (AMATYC). A common theme across these frameworks was the ability to solve mathematics problems in everyday situations using skills such as interpretation, reasoning, and representation (Sons 1996; Rhodes 2010; OECD 2012; Adelman et al. 2014). Existing assessments that measure quantitative skills in higher education (in addition to other general and subject-specific skills) included assessments such as the Collegiate Learning Assessment+ (CLA+), Collegiate Assessment of Academic Proficiency (CAAP), ETS Proficiency Profile (EPP), Graduate Record Exam (GRE), and the College-Level Examination Program (CLEP) to name a few. Roohr et al. (2014) also

reviewed existing item types, item formats, assessment structure, and content assessed. The majority of these assessments were multiple-choice assessments administered on a computer.

Assessment Framework

After reviewing the existing frameworks, definitions, and assessments, Roohr et al. (2014) identified the specific knowledge and skills to be assessed on *HEIghten* Quantitative Literacy by developing a theoretical assessment framework. This framework focused on two key areas: *problem-solving skills* and *mathematical content*. Primary problem-solving skills involved a student's ability to demonstrate skills in (a) *interpretation* of mathematical terms and representational devices; (b) *strategic knowledge and reasoning* to build, develop, and validate mathematical strategies, test conjectures, and draw appropriate inferences and conclusions; (c) *modeling* information into mathematical forms and applying and revising those models as needed; and (d) *communication* of mathematical concepts, data, procedures and solutions in a variety of forms. Students are also expected to be able to demonstrate computation skills to solve mathematical problems. Mathematical content included (a) number and operations, (b) algebra, (c) geometry and measurement, and (d) statistics and probability. These problem-solving skills and content areas work together. That is, students may have to *communicate* mathematical information using *statistics and probability.* See Roohr et al. (2014) for a deeper description of both the problem-solving skills and mathematical content areas.

Item Formats and Task Types

After developing the theoretical framework, the next steps involved identifying item formats and task types to measure these aspects of quantitative literacy. Roohr et al. (2014) identified numerous item formats and task types that could be used when developing a next-generation quantitative literacy assessment. When utilizing this framework to guide assessment development, item formats were selected based on their ability to ensure accurate construct coverage and accessibility for all students. For instance, although an open-ended graph item (i.e., an item that requires the examinee to graph the result instead of selecting the result from options) may be more authentic, this item format poses accessibility concerns. For a visually impaired student, this item format would need to be administered very differently. One possibility would be to provide tactile graphic materials, which could be hand scored. Another alternative could be the use of haptic technology (e.g., mechanical simulations such as vibrations when touching a tablet or smart phone). However, in both cases additional research would be needed before these item formats could be used operationally.

After considering construct coverage and accessibility, item formats for the pilot forms included single-selection and multiple-selection[3] multiple-choice, numeric entry, fraction entry, and grid[4] items. All stimuli for test items were embedded in real-world contexts including personal and everyday life, the workplace, and society. Additionally, some stimuli included word problems supplemented with accessible graphs, tables, or figures. Some questions also asked examinees to compare the relationships between two quantities. A four-function calculator was permitted to reduce the computational load on the test items, allowing for more of a focus on the problem-solving skills.

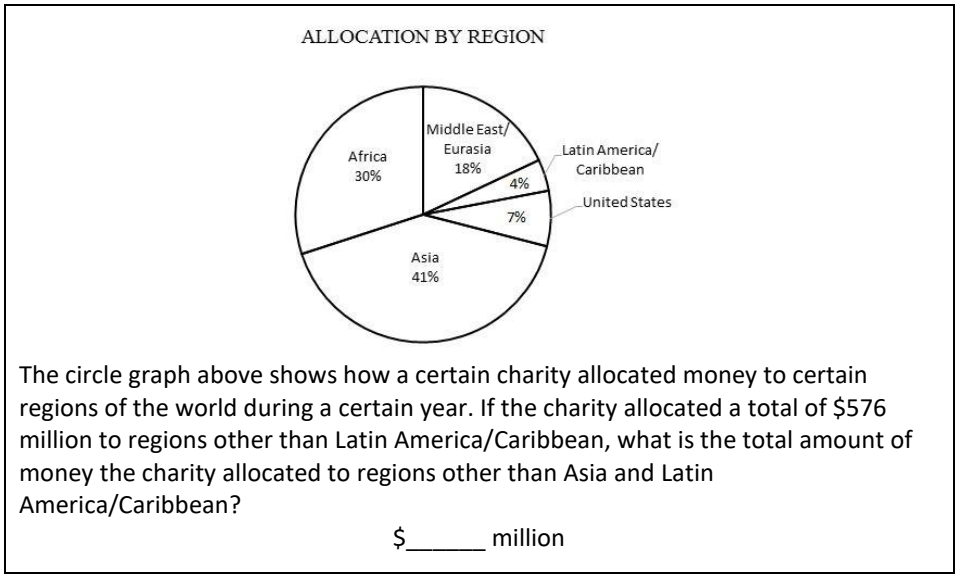

Figure 1. Sample Quantitative Literacy Item. It is a numeric entry item that measures Interpretation (problem-solving skill) and Number and Operations (content area). Note: the sample item is for reference only. It is not an actual question currently used on the assessment (ETS n.d.).

Each test item on the *HEIghten* Quantitative Literacy assessment was developed utilizing the research-driven framework. Test items measure at least one problem-solving area and one content area, and they are embedded into a real-world context. Figure 1 provides a sample test item. This numeric entry item measures the problem-solving skill of interpretation in the content area Number

[3] Multiple-selection multiple-choice items are selected-response items with multiple answer choices where one or more could be a correct response. For this assessment, these items were scored dichotomously (i.e., an examinee had to select all correct responses to get the item right).

[4] Grid items include a table with statements where the correct property is selected by check-marking a cell in the table. For this assessment, these items were scored dichotomously (i.e., an examinee had to select all correct responses to get the item right).

and Operations. Because test items were based on the comprehensive assessment framework, this framework served as validity evidence based on test content.

Study Purpose, Research Questions, and Rationale

Test design and development is an iterative process involving the adjustment and modification of the test items in response to data from testing out those items, as discussed in the *Standards for Educational and Psychological Testing* (AERA et al. 2014). The procedures around test design and development should support the validity of test scores. Focusing on the third step of test development, the purpose of this study was to examine the psychometric quality of the pilot items, evaluate individual- and group-level reliability, and provide preliminary validity evidence for *HEIghten* Quantitative Literacy, a next-generation assessment in higher education. Specifically, we addressed the following research questions:

1. How difficult and discriminating are the pilot test items? Are there some items that should be removed prior to the development of operational test forms?
2. What is the relationship between performance on problem-solving skill and content sub-areas?
3. What are the institution- and student-level reliability estimates?
4. How are the scores related to other variables such as high school and college grade point average (GPA), number of mathematical college courses taken, self-rated quantitative literacy skills, and college admissions scores?
5. What are the student perceptions of the assessment and how are their perceptions related to their test performance?
6. What are the differences in performance across various college-related subgroups (institution type, credit hours, and college major) and demographic subgroups (gender, race/ethnicity, language)?

All analyses involved data from a pilot test involving both two- and four-year higher education institutions across the United States. Results from this study will help support the development and use of this assessment operationally. For instance, item-level results will help determine which items to use when assembling operational test forms. Reliability estimates will inform us whether we have adequate consistency of test scores, and validity evidence will help support whether we are measuring the construct we intend to measure. Additionally, evaluating the relationships with other variables and student perceptions, and evaluating subgroup differences can help to provide useful information to institutions and other stakeholders about student learning in quantitative literacy. For instance, lack of learning gains or relationships with the number of mathematics courses taken in college may point to the need to emphasize quantitative literacy in the general education curriculum. Differences in college

majors can also provide some insight into the need to refocus the curriculum. These implications will be considered when interpreting these results.

There are some other well-known quantitative literacy instruments where similar studies have been conducted, including the Quantitative Literacy/ Reasoning Assessment (QLRA) instrument (Gaze et al. 2014) and the Quantitative Reasoning Test, Version 9 (QR-9; Sundre 2008). Similar to *HEIghten*, the QLRA is intended to be used across multiple campuses, and similar to the QR-9, *HEIghten* is intended to measure quantitative literacy skills that students may have obtained through a general education curriculum. This study goes beyond these existing studies by also capturing information on student performance across student subgroups, various college-related variables, and student perceptions. Evaluating performance across subgroups (e.g., college major) can provide institutions with information about which subgroups of students may be struggling in terms of their quantitative literacy skills. Additionally, the advantage of using an assessment such as *HEIghten* is that institutions can directly compare their performance to other institutions using the assessment. It is a national assessment that allows for benchmarking.

Method

Data and Sample

Data for this study included pilot data collected in March and April 2015 from 1,532 undergraduates across 23 institutions in the United States. Institutions volunteered to participate in this pilot study and were responsible for recruiting students within their respective institution. Institutions could use a number of incentives for recruiting students such as extra credit, course requirements, and financial incentives. Although there was some variation in how institutions recruited individual students to participate in the study, this variation is common practice when higher education institutions administer SLO assessments.

In total, this study included 438 students from seven two-year colleges, and 1,094 students from 16 four-year colleges. These institutions were fairly representative of institutions typically administering these standardized SLO assessments in terms of demographic breakdown. As shown in Table 1, the majority of students were female (61%), spoke English as a first language (85%), and were White (56%). More than half of the students were either freshmen or sophomores as indicated by the number of credit hours completed. Including students throughout their college career was important due to the number of different ways in which institutions can use these SLO assessments. For instance, some institutions might administer these assessments to a cohort of freshmen and a cohort of seniors then use these two cohorts to evaluate student learning gains from freshman to senior year cross-sectionally.

Table 1.
Sample Demographics

Demographic Information	N	Percent[a]
Institution Type		
2-year institution	438	29%
4-year institution	1094	71%
Gender		
Male	559	36%
Female	941	61%
Other/Missing	32	2%
First Language		
English	1296	85%
Other languages	197	13%
Missing	39	3%
Race/Ethnicity		
American Indian/Alaskan Native	11	<1%
African American/Black	261	17%
Asian/Asian American	128	8%
Hispanic/Latino	100	7%
White	858	56%
Multirace/Other	117	8%
Missing	57	4%

[a]Numbers may add up to slightly less or more than 100% due to rounding.

Instrument

As part of the pilot study, six *HEIghten* Quantitative Literacy test forms were administered on a computer to students using a spiraling approach to provide randomly equivalent distribution of students across test forms (i.e., when each student sat down at the computer to take the assessment, students were randomly assigned to test forms within and across the institutions to ensure that groups were randomly equivalent across test forms). Approximately 250 students took each test form and were fairly equivalent in terms of background variables (e.g., gender, race/ethnicity, class status). Each test form was composed of 25 dichotomously scored items (i.e., items scored as right or wrong) and took 45-minutes to complete. Because this was a pilot test, these six test forms were not developed to be comparable; thus there were some differences in the level of difficulty across test forms. As a result, equating (see section below) was necessary to adjust for the differences in difficulty across the six test forms. That said, the forms were developed to be fairly comparable in terms of construct coverage.

In addition to the assessment, a background information questionnaire and posttest survey were administered to each student. The background questionnaire asked students to provide information about their demographic (e.g., gender, race/ethnicity) and academic (e.g., grade point average [GPA], college admissions scores) background. Additionally, the posttest survey asked questions about the

reasons for taking the assessment (e.g., course requirement, extra credit), self-rated quantitative literacy skills, the number of mathematics courses taken in college, and their perceptions of the assessment.

Analyses

Student motivation. Prior to analyzing the pilot data, we first conducted motivational screening to identify students who did not try hard on the exam. At the test level, we removed students who did not complete at least 75% of the assessment. Using this criterion, a total of 33 students were removed from the remaining analyses. Additionally, when conducting the item analyses at the individual-item level, we also removed students who rapidly guessed on items (i.e., responded to the item in 3 seconds or less) to obtain a more accurate estimate of item difficulty and discrimination.

Item analyses. Item analyses included evaluating item difficulty and item discrimination. Because there were six test forms, each with 25 items, we evaluated a total of 150 test items. Item analyses were conducted separately by test form. Item difficulty was evaluated by calculating the proportion of examinees who got the item correct. A value closer to zero indicated a very difficult item (i.e., a smaller proportion of students got the item correct), and a value closer to one indicated an easier item (i.e., a larger proportion of students got the item correct). As part of item difficulty, we also conducted distractor analyses to evaluate the proportion of examinees who selected the alternative responses. Distractor analyses can help identify potential mis-keyed items. It can also be used to identify plausible and implausible distractors.

Item discrimination was evaluated using point-biserial correlations (i.e., the correlation between a right or wrong response and the total test score) to see whether an item discriminated between high- and low-performing students (Allen and Yen 1979). Negatively discriminating items should be dropped from an assessment as negative values suggest that more low-performing examinees are answering the item correctly than are high-performing examinees (Allen and Yen 1979).

Correlations across sub-areas. Because the assessment was designed to measure problem-solving skills in different content areas, we looked at the relationships between the raw scores across the three problem-solving skills (Communication/ Interpretation;[5] Modeling; and Strategic Knowledge and Reasoning) and the relationships between the raw scores across the four content areas (Number and Operations; Algebra; Geometry and Measurement; and Statistics and Probability).

[5] Communication and Interpretation were combined into one problem-solving skill area because of the small number of Communication items.

These correlations can inform us about the distinctiveness between these potential sub-areas and can provide some preliminary insight about the dimensionality of the assessment. If the correlations across the sub-areas are high, then students who score high on Number and Operations, for example, would also be likely to score high on Statistics and Probability. Therefore, reporting separate scores for those two sub-areas may not be as meaningful, because the two sub-areas are not providing distinct information back to the institution or examinee.

When calculating the correlations, a correction for attenuation (i.e., true-score correlation) was used to account for measurement error in the test scores:

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \quad (1)$$

where $r_{x'y'}$ is the true-score correlation, r_{xy} is the correlation between the two sub-areas, and r_{xx} and r_{yy} are the individual-level reliability estimates for each sub-area. As seen in Equation 1, this correction uses reliability estimates to correct for measurement error and indicates what the correlation would be if each sub-area had perfect reliability. As a result, the higher the reliability estimates across each sub-area, the less the portion of observed variance is due to errors in measurement.

Equating the six forms. Equating is used to adjust for differences in difficulty across multiple test forms, allowing the forms to be used interchangeably (Kolen 1988). A random equivalent-groups design and circle-arc equating method (see Livingston and Kim 2009) were used to equate the six forms before conducting further analyses. The equated test scores were used for the remaining analyses (except when calculating individual-level reliability across test forms). The circle-arc equating function was selected because of its ability to handle small sample sizes with a small standard error of equating.

Test reliability. Individual-level test reliability was calculated for the total score and sub-areas across all students using coefficient alpha for each of the six test forms. For students, there are generally minimal consequences based on the scores of these assessments (e.g., scores on the assessment are not likely used in decision making for students, or in high-stakes decisions like graduation or admissions), and as a result they are low-stakes for students. Given these low stakes, reliability estimates that are considered satisfactory can be lower than those reported for high-stakes assessments.

In addition to the individual-level reliability, group-level or institution-level reliability was also calculated. Institution-level reliability reflects the consistency of institutional mean scores across repeated test administrations with another six test forms and sample of students at the same institutions. At the group level, the scores for the assessment have moderate to high stakes because scores can be used for accreditation and accountability purposes. As a result, higher reliabilities

than what is reported at the individual level are needed given the higher stakes of the assessment.

To calculate institution-level reliability a multi-step procedure was used. Using this procedure, we first calculated the mean performance at each school on each test form. Given that we had 23 institutions in this study, we then had a vector of means for each of the 23 institutions for each test form. Using these vectors of institutional means, we then calculated the correlations across all possible pairs of test forms. With six test forms, there were 15 possible pairs of forms. For each of the 15 correlations, we applied the Spearman-Brown formula with $k = 6$. This calculation adjusted for the fact that the school means in this study are based on six times as many items and six times as many students. The institution-level reliability was calculated using the mean of the 15 correlations.

Relationships to other variables. Validity evidence based on the relationship to other variables was evaluated by examining the relationship between quantitative literacy performance (i.e., proportion correct on a scale from 0 to 100) and high school GPA, current GPA, number of mathematics courses taken in college, self-rated quantitative literacy skills, and college admissions scores (i.e., SAT or ACT score). Separate one-way analyses of variance (ANOVAs) were applied to test the performance differences across the different groupings of students for GPA, college classes, and self-rated skills. Prior to each analysis, we tested for homogeneity of variance using the Levene statistic. If this statistic was statistically significant (meaning that variances were unequal) we reported results using Welch's *t*-test or Welch's ANOVA test.

To evaluate the effect size for each ANOVA, we used omega-squared (ω^2), where 0.01 is a small effect, 0.06 is a medium effect, and 0.14 is a large effect. If the ANOVA was statistically significant, we conducted post-hoc analyses to evaluate individual differences between groups using the Bonferroni or Games-Howell (if using the Welch test) correction for family-wise error and Cohen's *d* to evaluate effect sizes. Cohen (1988) indicated that effects (i.e., Cohen's *d*) of 0.20 are considered small, 0.50 is moderate, and 0.80 is large. Note, however, that even a small effect can be viewed as important depending on the theory being tested (Gall et al. 2007). To give these effects context, we will consider results from other studies such as Liu et al. (2016) who conducted ANOVAs to evaluate performance differences on the *HEIghten* Critical Thinking measure.

Pearson correlations were conducted to evaluate the relationship between student quantitative literacy performance and college admissions scores. The guidelines by Cohen (1988) were used to evaluate the magnitude of the correlations, where 0.10 is small, 0.30 is moderate, and 0.50 is large. Correlations were conducted between quantitative literacy performance and SAT mathematics, critical reading, and writing, and with ACT mathematics, science, English, and reading. We also calculated the relationship between quantitative literacy

performance and SAT total (a composite score of critical reading and mathematics) or ACT total scores after converting ACT total scores into SAT total using the SAT-ACT concordance table (ACT 2013). If students reported both SAT and ACT score, their SAT score was used for this analyses. For the SAT/ACT total score we also calculated the disattenuated correlation to adjust for measurement error in the SAT/ACT and quantitative literacy scores.

Although the SAT mathematics construct does not have complete overlap with the construct of quantitative literacy, there is some overlap. SAT mathematics does measure both traditional mathematics as well as quantitative reasoning, so there is a slight overlap in some of the skills assessed on both assessments. As a result, we hypothesized that there would be small to moderate relationships between the two assessments. We also hypothesized that there would be some relationship with SAT critical reading given that all of the items on *HEIghten* Quantitative Literacy are word problems embedded in real-world contexts, so the reading load is higher for this assessment.

Relationship with student perceptions. Validity evidence based on the relationship with student perceptions was evaluated by examining the relationship between quantitative literacy performance and student perceptions about test difficulty (i.e., too difficult, at the right level, and too easy) and testing time (i.e., not enough time, enough time, more than enough time) based on their responses to the posttest survey. To evaluate student's perceptions on test difficulty and testing time, two separate one-way ANOVAs were conducted using ω^2 to evaluate the magnitude of the differences and conducting post-hoc analyses when appropriate. Students were also asked whether they tried their best when taking the assessment. To evaluate whether there were any statistically significant differences, we calculated an independent-samples *t*-test using Cohen's *d* to evaluate the magnitude of the differences.

Subgroup differences. Subgroup differences were evaluated for the following college-related variables: (a) institution type (2-year vs. 4-year), (b) credit hours (< 30, 30-60, 61-90, and > 90 credit hours), and (c) college major categories (business, natural science, humanities, and social science). Institution-type differences were evaluated using independent samples *t*-tests, and credit-hours and college-major differences were evaluated using one-way ANOVAs and one-way ANCOVAs with college admissions score as a covariate to control for prior achievement.

Subgroup differences were also evaluated for the following demographic groups: (a) gender (male vs. female), (b) race/ethnicity (Asian or Asian American, Black or African American, Hispanic or Latino, White, and Other), and (c) language (English speaking vs. non-English speaking). Gender and language differences were evaluated using independent-samples *t*-tests, and racial/ethnic

differences were evaluated using a one-way ANOVA. For all analyses, the magnitude of differences were evaluated using omega-squared (ω^2; for one-way ANOVA), or Cohen's *d* (*t*-tests and post-hoc tests).

Results

Item Analyses

Item difficulty. A total of 150 items were administered across the six pilot test forms. Item analyses revealed a mean proportion correct value of 0.36 across test items (standard deviation (SD) = 0.22) (Table 2), with values ranging from 0 to 0.89. Existing SLO assessments such as the CLA+ and CAAP typically aim for an item difficulty range of 0.30 to 0.80 (ACT 2012; Council for Aid to Education 2015), so our data indicated that we had some items with very low item difficulty as compared to existing SLO assessments. Across test forms, mean item difficulty ranged from 0.31 to 0.39 (Fig. 2). It is important to note that because this study was a pilot administration, test forms were not developed to be equivalent because we did not have any information on item difficulty prior to the study. As a result, there was a range of difficulty across the test forms. This circumstance was the reason that we conducted the circle-arc equating to adjust for differences in the difficulty across forms prior to conducting additional analyses in this study.

Table 2.

Item Difficulty and Discrimination

	Total Nbr of Items	Mean Item Difficulty	Mean Item Discrimination
Overall	150	.36	.41
Content Area			
Number & Operations	50	.40	.41
Algebra	35	.28	.38
Geometry & Measurement	33	.36	.42
Statistics & Probability	32	.34	.43
Problem-solving Skill Area			
Communication	5	.53	.36
Interpretation	44	.36	.42
Modeling	53	.33	.42
Strategic Knowledge & Reasoning	48	.36	.39
Item Type			
Single-Selection Multiple Choice	54	.46	.41
Quantitative Comparison	44	.38	.38
Multiple-Selection Multiple Choice	12	.18	.38
Numeric Entry	31	.23	.47
Fraction Entry	3	.20	.40
Grid	6	.22	.36

Note. Item difficulty = mean proportion correct; item discrimination = mean point -biserial correlations.

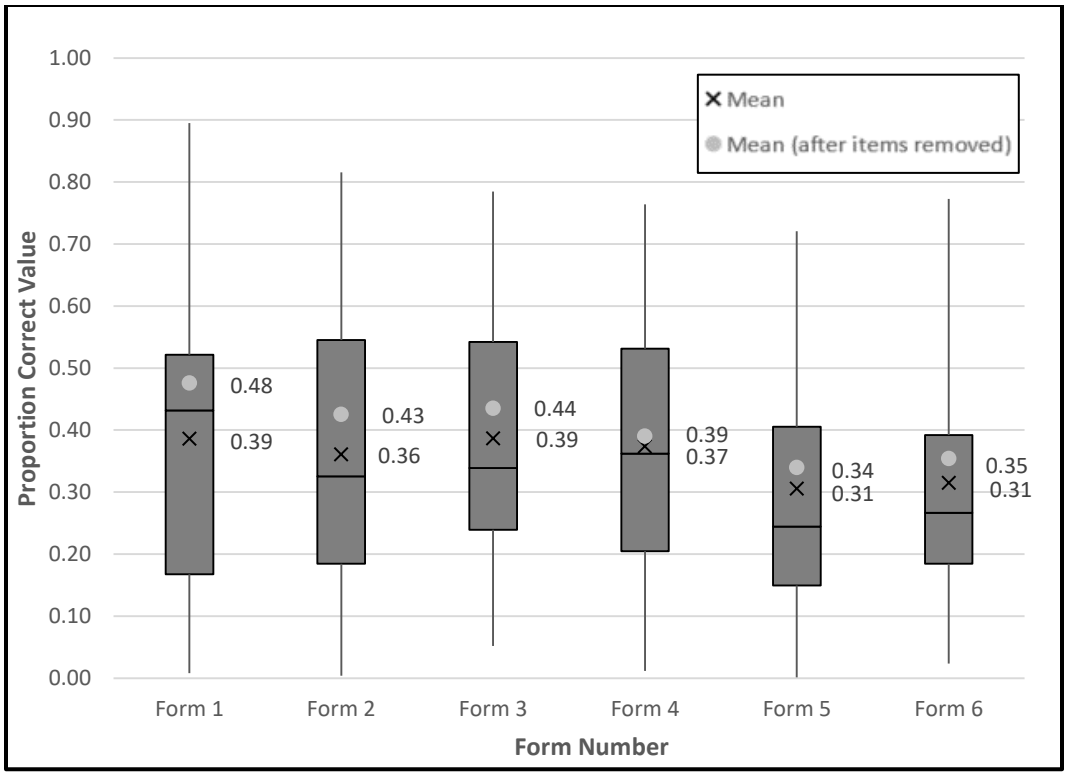

Figure 2. Item-Difficulty Distributions Across the Six Test Forms Prior to Equating. Note that these test forms were not created to be equivalent in terms of item difficulty because this administration was the first time testing these items. The box-plots show the mean item difficulty of all 25 test items, and the mean item difficulty after some items were removed due to very low difficulty. Notice that the mean item difficulty was higher after removing these items. Equating these forms adjusted for differences in item difficulty across forms.

When examining item difficulty, a total of 44 items showed values less than 0.20, with 21 items showing values less than 0.10. These items were further investigated for mis-keys using distractor analyses. All items were keyed correctly, but the 21 items with very low proportion correct values (< 0 .10) were removed from further analyses. After removing these 21 items, the mean proportion correct value across test items was 0.40 (SD = 0.20) and ranged from 0.35 to 0.48 across the six test forms (Fig. 2). This assessment included various item formats other than traditional single-selection multiple-choice items. As a result, the threshold for removing items was lower because some of the items had a lower than random chance level. Most of the items that were removed were numeric entry, multiple-selection multiple-choice, and grid items.

Item difficulty varied across item content area and problem-solving skill area. On an individual test form, item total for each content area ranged from 5-10 items for Number and Operations; 3-8 for Algebra; 4-7 for Geometry and

Measurement; and 3-7 for Statistics and Probability. Across all test items, Number and Operations items showed the largest proportion of students answering the items correctly on average, and Algebra items showed the smallest proportion of students answering the items correctly on average (Table 2). For the problem-solving skill areas, all three areas except Communication showed relatively equal proportion correct values, which was likely due to the small number of test items measuring Communication.

Lastly, we also evaluated item difficulty across item types (see Table 2). Results showed that traditional single-selection multiple-choice items showed the largest proportion of students answering items correctly, followed by quantitative comparison items.[6] This result could be partially due to the fact that students had a 25% chance of getting these items correct even when guessing. Item formats such as multiple-selection multiple-choice, grid, and fraction/numeric entry generally showed fewer students answering those items correctly on average.

Item discrimination. Across all test items, point-biserial correlations ranged from –0.02 to 0.66 with a mean of 0.41 (SD = 0.14) (Table 2). This result indicated that the assessment overall had good discriminating test items and that students who did well on an item were more likely to do well on the test as a whole. Mean point-biserial correlations ranged from 0.38 to 0.47 across test forms. Only one item showed a negative point-biserial correlation. That same item had a proportion correct value of 0.05, meaning very few students answered the item correctly. Across content areas, problem-solving skill areas, and item types, the mean point-biserial correlations were fairly similar (Table 2).

Summary. Given the results of the item analyses, a selected number of items were dropped from subsequent analyses due to unsatisfactory performance. Analyses revealed 21 items with difficulty values lower than 0.10 (and one item displaying a negative biserial). As a result, these 21 items out of 150 items were removed.

Due to the removal of 21 items, the total number of items for each test form were slightly different. The total items remaining across Forms 1 through 6 were 20, 21, 22, 23, 20, and 22 items, respectively. Because of the variation in items, the proportion of correct items across each form (0 to 100%) was used to represent the quantitative literacy scores during the analyses. The proportion correct with the six new test forms was used to equate the scores (i.e., adjust for differences in item difficulty) across forms using the circle-arc equating method. These equated scores were used throughout the remaining analyses. After

[6] Quantitative Comparison items are single-selection multiple-choice items where an "examinee compares two presented quantities (less than, equal to, or greater than) or determines that there is not enough information to make a comparison" (Roohr et al. 2014, 21).

equating the six forms, mean performance for Forms 1 to 6 were 36%, 35%, 36%, 35%, 34%, and 34%, respectively.

Correlations across Sub-areas

Table 3 shows both the observed and true-score (disattenuated) correlations across the three problem-solving skill areas. Given that measurement error can result in lower observed correlation coefficients, we focused on the true-score correlations when evaluating these results. Results across the problem-solving skills showed that the true-score correlations were all 0.92 or more, indicating a strong relationship between the different skill areas.

Table 3.
Observed and True-Score (Disattenuated) Correlations across Problem-Solving Skill Areas

	CI with M		CI with S		M with S	
	Observed	True-Score	Observed	True-Score	Observed	True-Score
Form 1	.55	.99	.52	1.00	.41	.93
Form 2	.47	1.00[a]	.59	.95	.53	1.00
Form 3	.58	1.00	.50	1.00	.65	1.00
Form 4	.56	1.00	.46	1.00	.52	1.00
Form 5	.58	1.00	.56	1.00	.63	.96
Form 6	.43	.92	.49	.98	.47	.99

Note. [a]Some correlations were greater than 1.00.
CI=communication/interpretation; M=Modeling; S=Strategic Knowledge and Reasoning.

Table 4 shows the correlation results across the four content areas. Similar to the problem-solving skill correlations, true-score correlations were also very high between many of the four different areas. The true-score correlations between Number and Operations (NO) and Geometry and Measurement (GM) were all above 0.93. Similarly, the true-score correlation between GM and Statistics and Probability (SP) were all above 0.85. A few lower true-score correlations were found across individual test forms. For instance, for Form 5, a true-score correlation of 0.67 was found between Algebra (AL) and GM, and a true-score correlation of 0.54 was found between AL and SP.

Table 4.
Observed and True-Score (Disattenuated) Correlations across Mathematical Content Areas

	NO with AL		NO with GM		NO with SP		AL with GM		AL with SP		GM with SP	
	Obs	T-S	Obs	T-S	Obs	T-S	Obs	T-S	Obs	T-S	Obs	T-S
Form 1	.43	.89	.52	1.00	.36	.76	.39	.84	.32	.69	.39	.85
Form 2	.47	1.00[a]	.46	.95	.59	1.00	.37	.96	.40	.92	.47	.99
Form 3	.40	1.00	.59	1.00	.47	.82	.33	1.00	.23	.90	.46	.94
Form 4	.41	1.00	.37	1.00	.43	1.00	.39	.90	.54	1.00	.46	1.00
Form 5	.41	.80	.59	.93	.56	.93	.32	.67	.25	.54	.53	.95
Form 6	.43	1.00	.50	1.00	.33	1.00	.54	1.00	.31	1.00	.34	1.00

Note. [a]Some correlations were greater than 1.00.
NO=Number & Operations; AL=Algebra; GM=Geometry & Measurement; SP=Statistics & Probability;
Obs=observed, unadjusted correlations; T-S=true-score correlation (disattenuated correlation).

Overall, these results suggested that the various sub-areas were not very distinct (i.e., scores across sub-areas were highly correlated) across the test forms. These results indicate that students who performed high on one problem-solving skill or content area were likely to also perform high on another skill or content area. These results suggest that providing sub-area scores back to an individual or institution may not be that meaningful, and that providing a total score would likely be sufficient given the strong relationships across sub-areas.

Reliability

Individual-level reliability using coefficient alpha revealed reliability estimates for the total score ranging from 0.72 to 0.83 across the six test forms (Table 5). For the sub-areas, however, individual-level reliability estimates were very low. These low estimates support our decision to report only the total score results to the individual test-taker and not report the sub-area results.

Table 5.
Individual-Level (Coefficient α) and Institution-Level Reliability Estimates

				Content Area[a]				Problem-Solving Skill[a]		
	N	Nbr Items	Total	NO	AL	GM	SP	CI	M	S
Individual-Level Reliability										
Form 1	257	20	.75	.50	.47	.47	.45	.62	.50	.39
Form 2	256	21	.78	.55	.35	.35	.54	.61	.22	.63
Form 3	252	22	.80	.68	.14	.50	.48	.44	.68	.49
Form 4	256	23	.76	.31	.50	.38	.51	.43	.63	.37
Form 5	256	20	.83	.68	.39	.59	.54	.47	.69	.63
Form 6	255	22	.72	.41	.45	.47	.15	.50	.44	.51
Institution-Level Reliability										
All Forms	1530	128	.96	.94	.85	.88	.90	.94	.92	.91

Note. [a]We do not plan to report scores across content area and problem-solving skill area back to the individual test taker due to the very low reliability estimates. Sub-area scores will only be reported at the institution-level. Only total score will be reported back to the individual student.
NO=Number and Operations; AL=Algebra; GM=Geometry and Measurement; SP=Statistics and Probability; CI=communication/interpretation; M=Modeling; S=Strategic Knowledge and Reasoning.

These individual-level reliability estimates are fairly comparable to estimates reported on existing SLO assessments (e.g., EPP, CLA+, CAAP). For instance, for the EPP, the reliability estimate of the total score is 0.91 for the Standard form and 0.77 for the Abbreviated form with subscore reliabilities ranging from 0.68 to 0.84 (ETS 2010). Similarly, the CLA+ reports reliabilities of 0.81 for the total score, 0.77 for the performance task, and 0.76 for the selected-response questions with subscore reliability estimates ranging from 0.51 to 0.58 (Council for Aid to Education 2015), and the CAAP reliability estimates range from 0.84 to 0.92 across test forms (ACT 2012).

For institution-level reliability, the linked scores based on the circle-arc equating function were used. Results yielded reliability estimates of 0.96 for the total score. Across content area and problem-solving skills, institution-level

reliability estimates ranged from 0.85 to 0.94. These results suggest that if sub-area results were to be reported, it is appropriate to report at the group- or institution-level. These results are slightly higher than existing SLO assessments (e.g., EPP, CLA, and CAAP) where institution-level reliability has ranged from 0.68 to 0.95 for freshmen, and from 0.64 to 0.93 for seniors (Klein at al. 2009).

Relationships with other Variables

GPA. Relationships between quantitative literacy performance (i.e., the equated percent correct score on the scale of 0 to 100) and high school and cumulative college GPA showed that students with a higher GPA tended to score higher on the quantitative literacy assessment (Fig. 3). Because the Levene's test for homogeneity of variance was statistically significant, we conducted two one-way

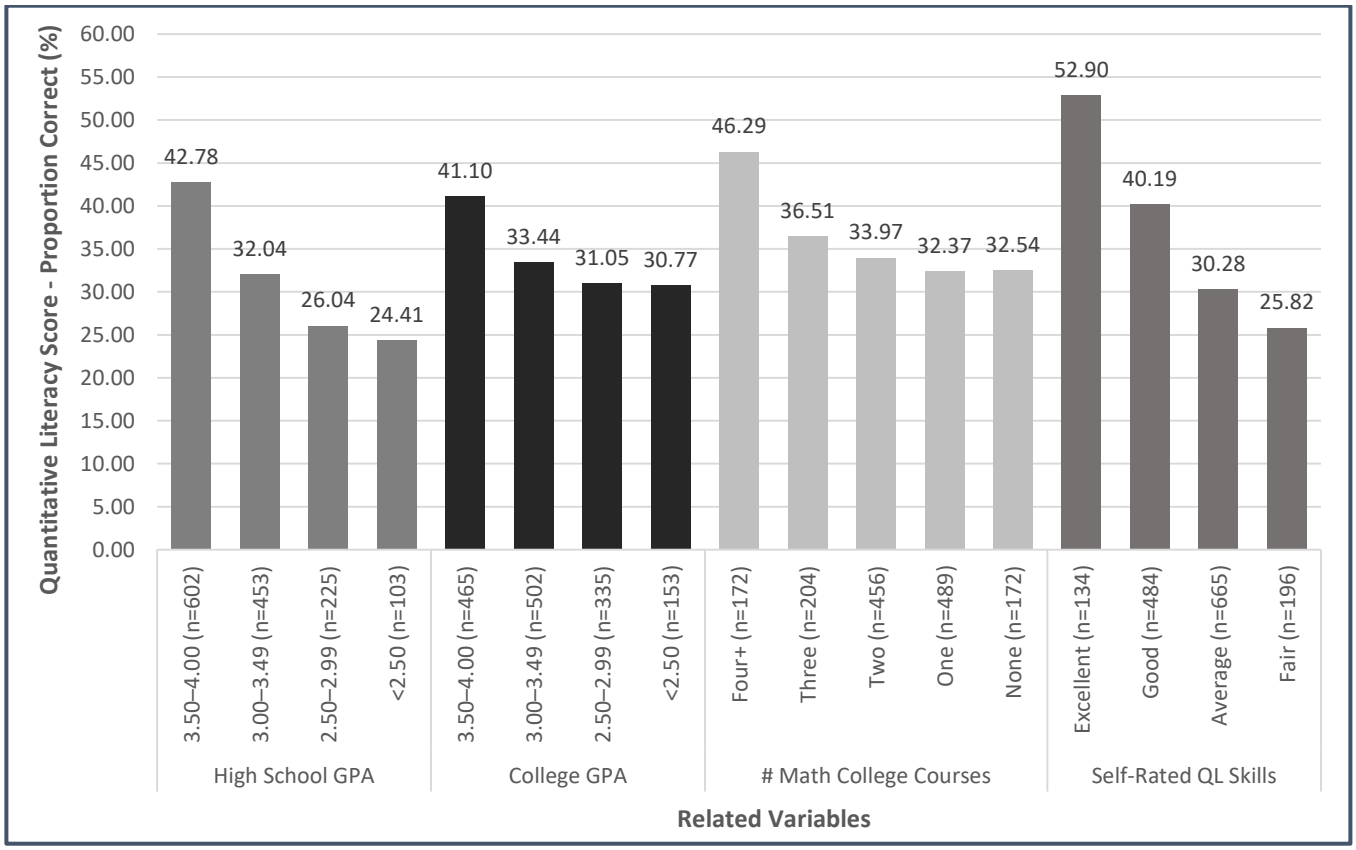

Figure 3. Quantitative Literacy Performance Across Related Variables (grade point average, number of mathematics-related courses in college, and self-rated quantitative literacy skills).

Welch ANOVAs.[7] Results revealed statistically significant differences in quantitative literacy score across the four categories for high school GPA (Welch's $F(3, 406.3) = 74.52$, $p < .001$, $\omega^2 = 0.14$) and cumulative college GPA (Welch's $F(3, 561.7) = 22.04$, $p < .001$, $\omega^2 = 0.04$).

For high school GPA, post-hoc analyses using the Games-Howell non-parametric tests revealed large statistically significant differences between the highest high school GPA category (3.50 to 4.00) and all other categories, with students in that category scoring 0.57 to 0.93 standard deviations (SDs) higher than students in the other categories (Fig. 3). Students in the second highest high school GPA category (3.00 to 3.49) also scored statistically significantly higher than the two bottom categories with moderate effect sizes ranging from 0.38 to 0.48. In terms of the magnitude of these differences, results were comparable to Liu et al. (2016). Similar results were found for college GPA. Students in the top college GPA category (3.50 to 4.00) scored statistically significantly higher than students in all other groups with moderate effect sizes ranging from 0.40 to 0.51.

Number of college mathematics classes. Results of the one-way Welch's ANOVA indicated a statistically significant difference in quantitative literacy score depending on the number of mathematics courses taken in college (Welch's $F(4, 527) = 13.78$, $p < .001$, $\omega^2 = 0.03$). Results showed that students who have taken more mathematics college-level classes performed higher on the quantitative literacy measure (Fig. 3). The only statistically significant differences in scores (Games-Howell post-hoc results) were with students who took four or more college-level classes as compared to all other students who took fewer classes. Specifically, students taking four or more classes showed moderate to large differences in performance ranging from 0.47 to 0.73 SDs as compared to their peers who took fewer classes.

Self-rated quantitative literacy skills. Fig. 3 also shows students' quantitative literacy performance in relation to their self-rated skills. Results showed that students who rated themselves with strong quantitative literacy skills typically performed higher than students who rated themselves with lower skills. One-way Welch's ANOVA results showed statistically significant differences in performance across the four self-rated skill categories with a very large effect (Welch's $F(3, 431.8) = 78.82$, $p < .001$, $\omega^2 = 0.14$). Post-hoc comparisons showed that all four groups were statistically significantly different from each other. For instance, students who rated themselves as having excellent quantitative literacy skills scored 0.60 to 1.45 SDs higher than all other students.

[7] For the remaining analyses, we simply reported the Welch results if the Levene's test for homogeneity of variance was significant.

SAT/ACT score. SAT or ACT scores were available for 58% of the sample of students who took the quantitative literacy assessment. To evaluate whether this sample of students was representative of all the students who took the quantitative literacy assessment, we first evaluated performance differences between those who reported college admissions score and those who did not. Results showed that students reporting college admissions score performed statistically significantly higher by approximately 10 percentage points than those students who did not report a college admissions score (Welch's $t(1530) = -10.57$, $p < .001$, $d = 0.94$). These results suggest that those students reporting college admissions score may have been higher performers, which should be considered when interpreting these results. That is, these findings may not be representative of the full sample of students and should be replicated with all students.

Table 6.

Relationships with College Admissions Scores

	N	Percent Reported	*r*	Disattenuated r^a
SAT or ACT	890	58.09	.65**	.69
SAT				
Critical Reading	651	42.49	.56**	.59
Mathematics	619	40.40	.63**	.67
Writing	514	33.55	.51**	.55
ACT				
English	384	25.07	.50**	.53
Mathematics	409	26.70	.53**	.57
Reading	393	25.65	.41**	.56
Science	369	24.09	.53**	.58

Note. N=1532; Percent Reported = the proportion of students who reported valid SAT/ACT scores; *r* = correlation between proportion correct score (equated across forms) and college admissions score.
[a]Calculated using the institution-level reliability of .96 for HEIghten.
**$p \leq .001$.

Correlations between quantitative literacy score and SAT/ACT total revealed a statistically significant positive correlation of 0.65 (Table 6). To adjust for any measurement error, we also evaluated the disattenuated correlations. The reliability of SAT scores is at least 0.91 (College Board, 2014), and the estimated institution-level reliability for the quantitative literacy score was 0.96. The disattenuated correlation was 0.69 between quantitative literacy score and SAT/ACT total. To give the magnitude of the correlations some context, we compared our results to similar studies investigating the relationship between SLO assessment score and college admissions score. For instance, Shavelson (2010) found correlations ranging from 0.55 to 0.57 between scores on the CLA and SAT. More recently, Liu et al. (2016) found a correlation of 0.54 between *HEIghten* Critical Thinking score and an SAT/ACT composite score. Those authors noted that the disattenuated correlations ranged from 0.63 to 0.71.

We also evaluated the relationship between quantitative literacy score and subscores on the SAT and ACT (Table 6). As expected, for SAT, results showed that quantitative literacy scores correlated highest with SAT mathematics (disattenuated $r(619) = 0.67$), followed by critical reading (disattenuated $r(651) = 0.59$) and writing (disattenuated $r(514) = 0.55$). For ACT, similar trends were found. Quantitative literacy scores correlated highest with ACT mathematics (disattenuated $r(409) = 0.57$) and ACT science (disattenuated $r(369) = 0.58$).

Relationships with Student Perceptions

Perceived test difficulty and testing time. The majority of students indicated that the assessment was on the right difficulty level (71%). One-way Welch's ANOVA results revealed statistically significant differences in performance across the three categories of perceived test difficulty (Welch's $F(2, 187.2) = 56.98$, $p < .001$, $\omega^2 = 0.07$). Students who perceived the test was too easy performed statistically significantly higher than all other students with very large effects ranging from 1.08 to 1.64 (Fig. 4). Students who felt as though the assessment was on the right level performed 0.39 SDs higher than students who felt the assessment was too difficult.

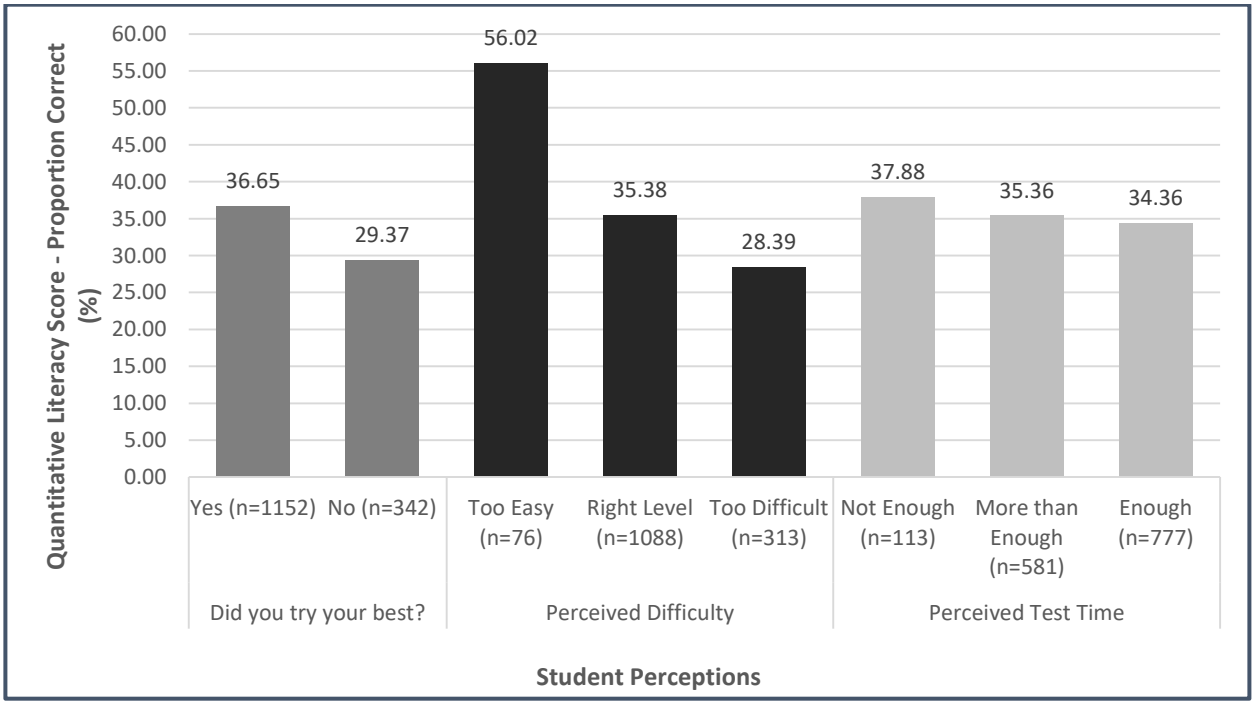

Figure 4. Quantitative Literacy Performance in Relation to Student Perceptions (student effort, perceived difficulty, and perceived testing time).

Additionally, most students indicated that they had either enough time (51%) or more than enough time (38%) to complete the assessment. Only 7% of students indicated that they did not have enough time to complete the assessment. Although students who perceived not to have enough time on the assessment performed the highest, one-way ANOVA results indicated that there were no statistically significant differences in performance across the three categories of perceived testing time ($F(2, 1468) = 1.84$, $p = 0.16$, $\omega^2 = .001$).

Self-reported effort. During the posttest survey, students were asked whether they tried their best when taking the assessment: 75% of students indicated yes ($n = 1152$); 22% indicated no ($n = 342$); and 2.5% did not respond. Students who indicated that they tried their best performed 7.18 percentage points higher than students who indicated that they did not try their best (36.65% compared to 29.47%), or by about 0.42 SDs; this performance difference was statistically significant (Welch's $t(635.1) = 6.77$, $p < .001$) (Fig. 4).

To further investigate student effort, we also looked at the 33 students who did not complete at least 75% of the assessment to see their self-reported effort. Of these 33 students, 30% indicated that they did not try their best when taking the test; 55% indicated yes, they did try their best; and 15% did not answer this survey question.

Subgroup Performance Differences

College-related subgroups. Differences in overall quantitative literacy performance were examined between institution type, credit hours, and college major (Fig. 5). Results indicated that students at four-year institutions performed approximately 8.4 percentage points higher than students at two-year institutions (Welch's $t(1057.3) = -8.94$, $p < .001$, $d = 0.45$).

In relation to college credit hours, differences across four categories were examined: (a) < 30 semester hours (i.e., freshmen), (b) 30-60 semester hours (i.e., sophomores), (c) 60-90 semester hours (i.e., juniors), and (d) > 90 semester hours (i.e., seniors). When running this particular set of analyses, we looked only at students at four-year institutions. One-way ANOVA results revealed no statistically significant differences across the four credit-hour categories ($F(3, 1075) = 0.08$, $p = 0.97$). Additionally, when controlling for prior achievement using college admissions scores, one-way ANCOVA results also revealed no statistically significant differences in performance across the four groups ($F(3, 714) = 1.69$, $p = 0.17$).

College majors were classified into four categories: (a) business, (b) humanities, (c) social sciences, and (d) natural sciences. Controlling for prior achievement using college admissions score, one-way ANCOVA results indicated statistically significant differences in performance across the four college-major classifications ($F(3, 447) = 5.97$, $p = .001$, $\omega^2 = 0.03$) with business majors

performing highest (45.6%) followed by natural sciences (39.0%), humanities (33.9%), and social sciences (28.4%). Post-hoc analyses using Bonferroni revealed statistically significant differences between business and natural science majors ($p < .001$, $d = 1.00$), and natural science and social science majors ($p < .001$, $d = 0.60$).

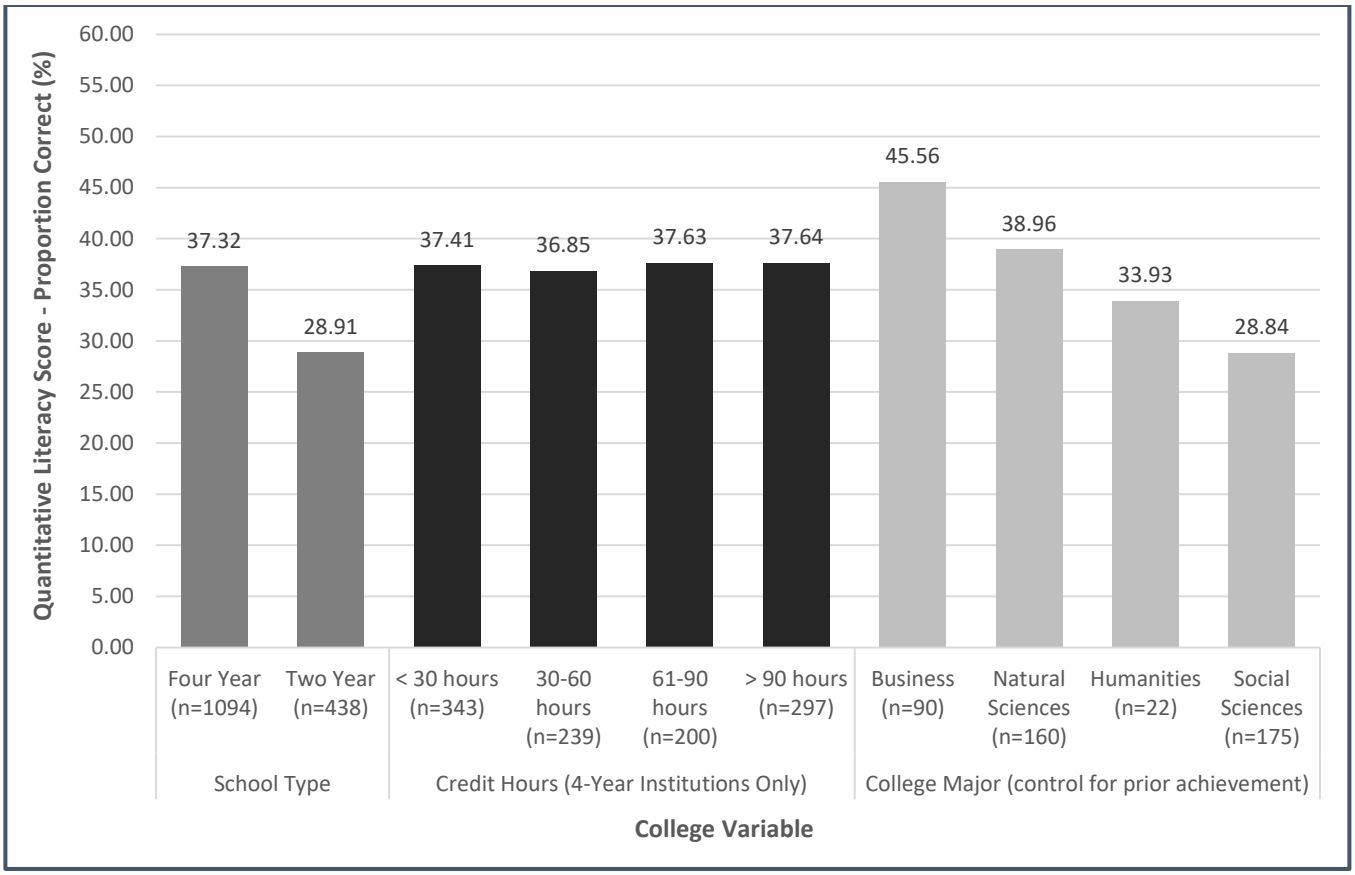

Figure 5. Quantitative Literacy Performance Across College Variables (school type, credit hours, and college major).

Demographic subgroups. Differences were also examined across demographic subgroups including gender, race/ethnicity, and language (Fig. 6). Results showed that males outperformed females by 7 percentage points (39% compared to 32%), or by about 0.36 SDs; this difference was statistically significant (Welch's $t(991.8) = 6.46$, $p < .001$). Previous studies found effect sizes ranging from zero to as much as 0.29 (Hyde et al. 1990; Lindberg et al. 2010; Liu and Roohr 2013). In comparison to these previous studies, the effect size here was slightly larger, indicating a larger difference between males and females.

Five different racial/ethnic groups were compared in terms of their quantitative literacy performance: (a) Asian or Asian American, (b) Black or

African American, (c) Hispanic or Latino, (d) White, and (e) Other. Welch's one-way ANOVA revealed large statistically significant differences across the five groups (Welch's $F(4, 327.7) = 55.40$, $p < .001$, $\omega^2 = 0.13$). Post-hoc analyses (Games-Howell) indicated that Asian/Asian American students performed statistically significantly higher ($p < .001$) than all other subgroups with effect sizes ranging from 0.84 to 1.66 SDs. Black/African American students performed statistically significantly lower than all other subgroups ($p < .001$, $d = –0.39$ to $–1.66$), and White students performed statistically significantly higher than Black/African American ($p < .001$, $d = 0.62$) and Hispanic/Latino students ($p < .01$, $d = 0.31$).

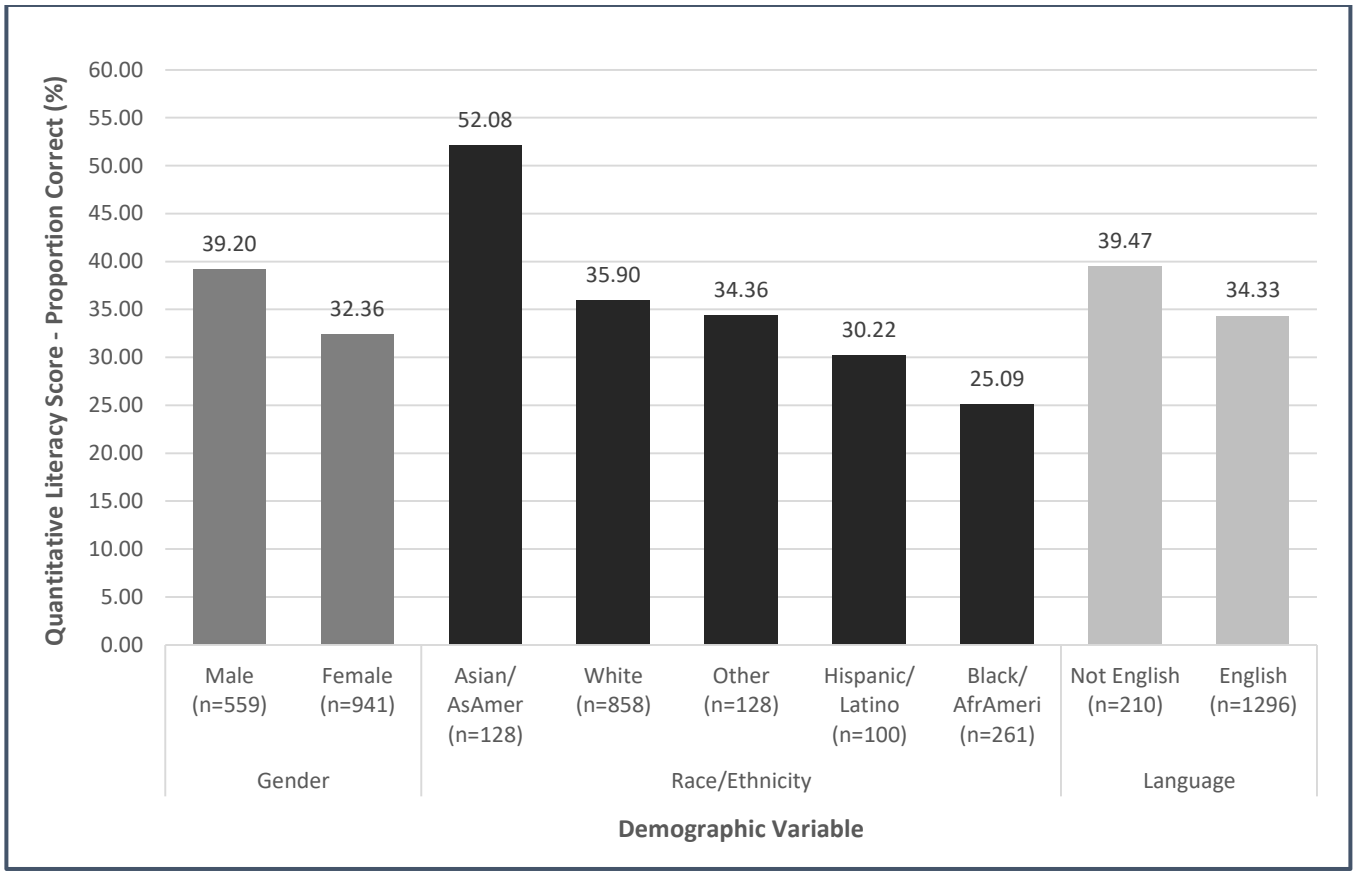

Figure 6. Quantitative Literacy Performance Across Demographic Variables (gender, race/ethnicity, and language).

Interestingly, results comparing students whose first/native language was not English versus students whose first/native language was English indicated that non-native English speakers performed statistically significantly higher by 5 percentage points or 0.27 SDs (Welch's $t(261.3) = –3.25$, $p < .001$). It is important to note, however, that the sample of non-native English speakers was

fairly low. We conducted a cross-tabulation to see which racial/ethnic groups comprised the non-native English speaker group. Results indicated that 36% self-identified as Asian/Asian American, 15% as White, and 11% as other, which could explain the performance differences.

Discussion

With the growing importance of quantitative literacy in higher education and in the workplace, it is essential that we evaluate whether students are developing these skills throughout college. *HEIghten* Quantitative Literacy is one measure that could be used to capture that information. The purpose of this study was to conduct a preliminary evaluation of the psychometric quality of the newly developed *HEIghten* Quantitative Literacy assessment by conducting item analyses, estimating individual and group-level reliability, providing preliminary validity evidence based on relationships to other variables and student perceptions, and evaluating subgroup differences using data from a pilot study.

Results from this study yielded the following conclusions: (a) overall, items functioned well;[8] (b) true-score correlations across sub-areas of the assessment were very high indicating that the assessment may be unidimensional; (c) reliability estimates similar to existing SLO assessments were found at the individual and group levels; (d) test scores showed positive relationships with high school and college GPA, number of college mathematics courses, self-rated quantitative literacy skills, and college admissions scores; (e) students reported positive perceptions about the assessment, and (f) performance differences were found across institution type, college majors, gender, racial/ethnic groups, and language groups, but not across credit hour categories. Results from this study provided preliminary validity evidence to support the use of *HEIghten* Quantitative Literacy at higher education institutions. Operational test forms will be developed based on the results from this pilot administration. All analyses will be replicated with operational data and additional analyses will be conducted to evaluate other sources of validity evidence.

Using the Results to Guide Operational Test Development

Results from the item analyses, correlations across sub-areas, and reliabilities can directly inform the development of operational test forms. Item analyses were able to reveal the psychometric properties of the test items showing that overall, items functioned well and many of the items can be used to assemble the operational test forms. Results showed that item difficulty ranged from 0.11 to

[8] Poorly performing items should be dropped when developing the operational test forms.

0.89 (after removing the 21 items with proportion correct values less than 0.10), and that items had good discrimination indices with a mean of 0.43. Item difficulty statistics were comparable to existing SLO assessments. For instance, the majority of the CAAP Mathematics items range in difficulty from 0.10 to 0.79 across the two test forms (ACT 2012). Moving forward, the following items will not be included operationally: items with item difficulty values less than 0.10, and negative point-biserials.

Test Unidimensionality

The theoretical construct for *HEIghten* Quantitative Literacy was multidimensional with both problem-solving skill areas (i.e., Interpretation; Strategic Knowledge and Reasoning; Modeling; Communication) and content areas (i.e., Number and Operations; Algebra; Geometry and Measurement; Statistics and Probability). However, high correlations across sub-areas revealed that the problem-solving skill and content areas were not very distinct, meaning that if we were to report subscores on these sub-areas, they would not provide meaningful information to the institution beyond the information provided by the total score.[9] These results suggested that students who performed high on one sub-area were likely to perform well on other sub-areas. The high correlations suggest that the assessment is practically unidimensional. That said, we will replicate these analyses (i.e., the correlation analyses) and also conduct factor analyses to determine the dimensionality of the final test forms developed based on the final content specifications. Although the assessment may be practically unidimensional based on these preliminary findings, providing subscore data at the group-level back to the institutions could potentially provide actionable data about areas of strength and weakness in quantitative literacy for students within the institution.

Results from this study are consistent with previous research on other mathematics assessments. For instance, the National Assessment for Educational Progress (NAEP) Mathematics has a multidimensional theoretical construct measuring five content areas: (a) numbers and operations, (b) measurement, (c) geometry, (d) statistics, and (e) algebra; however, empirical results have provided evidence for a unidimensional construct of mathematics (e.g., Rock 1991; Abedi 1997). For instance, results have shown that the five mathematics subscales were highly correlated with factor loading correlations ranging from 0.89 to 0.97 for Grade 12 (Abedi 1997). Similar results have been found on assessments for Grades 4 and 8 (Abedi 1997).

[9] It is important to note that due to low reliability estimates, subscores would not be provided to the individual test-taker, and instead would be provided only at the group- or institutional-level.

Similarly, the Programme for International Assessment (PISA) Mathematics also has a multidimensional theoretical framework focusing on three key areas: content, process, and context. Content areas include: quantity, space and shape, change and relationship, and uncertainty. Ekmekci (2013) investigated the dimensionality of the 2003, 2006, and 2009 PISA Mathematics assessments and found that although the multidimensional models fit the data well, correlations across the four content areas were very high (e.g., ranging from 0.91 to 0.99 for the 2003 data), thus providing evidence to support the unidimensional model. Although these previous studies are based on assessments in the K-12 space, these studies from NAEP and PISA demonstrate that there are other existing mathematics assessments that, despite being developed based on a multidimensional theoretical construct, have shown empirically to be a unidimensional construct. That is, results from previous studies have also shown that students who perform high on one sub-dimension are also likely to perform high on another, consistent with our findings in this study.

Positive Relationships with GPA and College Admissions Scores

Evaluating the relationship with other variables provided validity evidence to support that *HEIghten* Quantitative Literacy measures skills students had before entering college, and skills students have learned in college. Results showed that as high school and cumulative college GPA increased, quantitative literacy scores also increased, suggesting that pre-college and within-college academic performance has a relationship with students' quantitative literacy performance. These results are consistent with previous research also investigating the relationship between SLO performance and GPA (e.g., Kuncel et al. 2001; ACT 2012; Liu and Roohr 2013; Liu et al. 2016; Graduate Management Admissions Council 2017).

Results also showed positive relationships between college admissions scores and quantitative literacy scores with disattenuated correlations of 0.67 and 0.57 with SAT mathematics and ACT mathematics, respectively. These results provide evidence that performance on *HEIghten* Quantitative Literacy is related to performance on college admissions assessments. The magnitude of these correlations is fairly consistent with previous research with correlations ranging from 0.54 to 0.57 (see Shavelson 2010; Liu et al. 2016). Interestingly, ACT science was also correlated highly with quantitative literacy scores with a disattenuated correlation of 0.58. Lower correlations were found between quantitative literacy score and SAT critical reading and writing, and ACT English and writing. These results suggest that *HEIghten* Quantitative Literacy is in fact measuring a mathematics construct. It is important to note, however, that the *HEIghten* Quantitative Literacy is somewhat different because all test items were

embedded in real-world contexts, whereas the SAT and ACT mathematics sections may have more straightforward mathematics questions. Because of these contexts, there was a higher reading load on the quantitative literacy measure, which would explain why there were still large relationships with SAT critical reading and with ACT English and reading. It is also important to note that although there is a relationship between SLO assessment performance and college admissions scores, each assessment has a different purpose with the key difference being that SLO measures are intended to provide information that can help inform teaching and learning (Benjamin et al. 2009). That is, group-level performance on SLO assessments may be used by institutions to help gauge whether students are making learning gains from freshman to senior year. Institutions may also disaggregate group-level results by subgroups such as college major to see how students in groups are performing on various SLOs. These scores, along with student proficiency levels and proficiency level descriptors, can be used by institutions to help inform the general education curriculum within an institution.

The Impact of Mathematics Courses on Quantitative Literacy Performance

We also evaluated the relationship with the number of mathematics courses a student took in college. Results showed that students who took more mathematics courses in college typically performed higher on *HEIghten* Quantitative Literacy. In fact, students who had taken four or more classes performed statistically significantly higher than students taking three or fewer classes. It is interesting that students who took three courses did not perform statistically different from those taking two or fewer courses. Future research would benefit from knowing which mathematics courses students took and how they relate to performance. Quantitative literacy performance may be more related to the type of content that students are learning in various college courses, rather than the number of courses they take. For instance, Hughes-Hallett (2003) noted that many students take introductory college courses in mathematics, but fail to progress beyond the memorization of problem types. It may be that in order to demonstrate statistically significantly higher quantitative literacy performance, students need to take more than just introductory mathematics courses or courses that focus on rote memorization. In fact, Small (2003, 252) suggested that "the most effective way to advance quantitative literacy is to improve the traditional college algebra to serve as a foundation course for QL [quantitative literacy]." He suggested that these courses should focus on skills such as data analysis, modeling, developing communication skills, using appropriate technology, and participating in small group projects. Perhaps if there was a curriculum shift to focus more on these various skills, we would see improvements in students' quantitative literacy

performance. Using *HEIghten* Quantitative Literacy as an SLO measure can help provide information to institutions about where students are struggling and whether changes should be made to the general education curriculum to improve students' quantitative literacy skills.

Interestingly, despite showing a positive relationship between the number of mathematics college courses and quantitative literacy performance, overall quantitative literacy performance was quite low. This result was also found when examining relationships with other variables such as GPA. Additionally, when examining performance differences across completed credit hours, there were no significant differences across the four categories. Results showed that students with more than 90 completed credit hours did not perform statistically significantly different from students less than 30 completed credit hours. It could be that the majority of students are not taking mathematics courses in college that are contributing to their quantitative literacy skills. For instance, in our sample of students in this study, 42% either never took a mathematics course in college, or took only one course. Future research would benefit from looking into the general education course requirements at institutions. We should also consider working closely with institutions who have made changes in their curriculum to focus more on quantitative literacy skills. For instance, we could look at institutions using the Quantway program to see if these institutions who focus more on quantitative literacy perform higher than those without quantitative literacy-specific coursework. We could also evaluate whether enrolling in courses specifically targeted at quantitative literacy skills result in learning gains.

Student Effort

For this study, student effort was investigated using responses from an item on the posttest survey that asked students whether they tried their best on the assessment. Self-reported results indicated that 75% of students tried their best, which meant the majority of students likely put forth their best effort. These results are positive given that low motivation in low-stakes SLO assessment has been an area of concern (Klein et al. 2009; Liu 2011; Liu et al. 2012). Results showed a moderate and statistically significant difference in performance (0.42 SDs) between students who indicated that they tried their best and students who indicated that they did not try their best. To put this finding into context, this difference in performance matches the performance differences typically found between freshmen and seniors (Blaich and Wise 2011; Arum and Roksa 2014). Previous research has found similar differences in performance between motivated and unmotivated examinees (e.g., Wise and Kong 2005; Wise and DeMars 2010; Liu et al. 2012; Rios et al. 2014; Liu et al. 2015). These results stress the importance of considering student effort or motivation on low-stakes SLO measures.

Student perception results about the amount of testing time can also provide some insight into student effort. Although results indicated that there were no statistically significant differences in performance, the 7% of students who indicated that they did not have enough time on the assessment scored 2.5 percentage points higher than those who indicated that they had more than enough time, and 3.5 percentage points higher than those students who indicated they had enough time. It is likely that the students who perceived that they did not have enough time were actually those students with higher motivation levels. In fact, upon further analyses, results showed that of the 114 students who indicated perceived that they did not have enough time, 75% of them indicated that they tried their best, which could suggest some relationship between testing time and student effort. Liu et al. (2015) found that on average, motivated students spent 15 seconds longer on individual test items as compared to unmotivated students. Because higher motivated students may take longer on test items, they may have also felt as though there wasn't enough time to complete the assessment items.

Future research should further investigate the issue of student effort including using more methods than one self-report question. Self-report may not be an accurate way to capture student effort. Results showed that of the 33 students who did not complete at least 75% of the assessment, 55% still indicated that they tried their best. Previous research has shown that response time may be an effective way for detecting student effort (e.g., Wise 2006; Wise and Ma 2012). For this study, we did use 3 second rule to identify unmotivated students across individual test items; however, response time can be used to determine overall motivation across an assessment, and there are a number of different methods that can be used and considered for future research. Because the assessment is on the computer, we can easily collect individual response data. The method for detecting rapid responses is a much larger research question and will be further investigated in future research. We should also investigate how response time information can be used with self-report data to effectively identify motivated and unmotivated students. Additionally, future research should also evaluate methods to improve student motivation at the start of the assessment through methods such as changes in the instructions for the assessment, which have been found to be an effective method for motivating students (see Liu et al. 2012; Liu et al. 2015).

Performance Differences across Subgroups

Institution-type. Results revealed that students at four-year institutions performed significantly higher than students at two-year institutions. These results are consistent with previous research (Baer et al. 2006; Liu and Roohr 2013). For instance, results from the National Survey of America's College Students (NSACS), which used the National Assessment of Adult Literacy (NAAL) to measure quantitative literacy, also showed students at four-year institutions

outperforming their peers at two-year institutions (Baer et al. 2006). The NSACS had approximately 1,800 graduating students at 80 randomly selected two- and four-year institutions as part of their sample. Results from this study also showed that in general, students struggled the most with quantitative literacy as compared to prose and document literacy.

There are a few different reasons why we might be seeing performance differences across institution-type. One reason could be that the population of students enrolled in two-year institutions is inherently different from those students attending four-year institutions. For instance, students at two-year institutions are typically slightly older, are more likely to be the first generation to attend college, and are more likely to be working a job (American Association of Community Colleges [AACC] 2009). These students may also be enrolled in only a single course at the institution to update a specific job skill or earn a promotion (AACC 2009), making it difficult to measure what knowledge and skills the students are learning specifically at the institution as a result of the courses and activities they are engaged in (Nunley et al. 2011). Future research would benefit from conducting case studies at two-year institutions to better inform these performance differences. Knowledge of the curriculum in relation to required quantitative courses, and more background information about students enrolled in the institution would help to inform why we are seeing lower performance on average by students in two-year institutions.

College major. Results also showed statistically significant differences in performance across college majors when controlling for prior achievement using college admissions scores. Not surprisingly, business and natural science majors performed the highest and were not statistically different in terms of performance. Given that students in these major categories are more likely to enroll in quantitative courses, we would suspect their performance to be higher as compared to their peers in humanities and social science majors. These results point to the importance of improving quantitative literacy skills for students enrolled in a humanities or social science major. These results suggest the need to include more general education courses that are focused on quantitative literacy skills so that all students, regardless of college major, learn the appropriate skills upon graduating college and are prepared to enter the workforce community.

Gender. Results revealed a gender performance gap favoring males of 0.36 SDs, which is larger than results from previous research that have also evaluated gender performance differences (Hyde et al. 1990; Lindberg et al. 2010; Liu and Roohr 2013). Other research such as the NSACS, however, has shown that males and females did not perform significantly different in quantitative literacy at both two- and four-year institutions (Baer et al. 2006). After further evaluating the gender difference for this study, we found that the difference in performance

across gender may also be directly related to college major. Specifically, approximately 47% of the male sample were completing business or natural science majors compared to 37% of the female sample. Given that business and natural science majors performed statistically significantly higher, this could partially explain why we are seeing a gender difference.

Race/ethnicity. Results showed that Asian/Asian American students performed statistically significantly higher than all other subgroups and that Black/African American students performed statistically significantly lower than all other subgroups. White students performed the next highest followed by Hispanic/Latino students. These results are similar to trends found in K-12 national assessment results using NAEP (U.S. Department of Education 2014). Trends from K-12 are likely to remain as students enter college. That said, in higher education, there have been mixed results in terms of the differences between Asian and White students. For instance, PIAAC numeracy results found no significant differences in performance between Asian and White students (Goodman et al. 2013). Additionally, the NSACS showed White students significantly outperforming their Asian/Pacific Islander peers on quantitative literacy (Baer et al. 2006). These mixed results point to the need to further disaggregate the Asian/Pacific Islander subgroup to further understand the performance difference on quantitative literacy. It is also important to further understand ways to reduce the Black-White and Hispanic-White performance gaps in higher education.

Limitations and Future Research

Due to limitations in the data, we were unable to evaluate other sources of validity evidence. For instance, future research should further investigate the dimensionality of the assessment using confirmatory factor analysis. Because not all forms were comparable in terms of item difficulty, and given that slight adjustments will be made to the operational test forms, we plan to conduct these analyses on the operational forms that are appropriately balanced in terms of test content, item difficulty, and item discrimination. These analyses will allow us to further investigate whether we are measuring one dimension of quantitative literacy, or if the assessment is more multidimensional, capturing multiple dimensions such as content area (e.g., statistics and probability) or problem-solving skills (e.g., interpretation). Additionally, we should evaluate direct evidence of response processes using methods such as cognitive interviews or think-aloud procedures. Evidence should also be evaluated regarding the consequences of testing. That is, we should evaluate how institutions are actually using the assessments and evaluate potential unintended consequences and their impact on the interpretation of test scores. Lastly, all analyses from this study

should be replicated with the operational data to support the intended uses of test scores.

Another limitation to this study was the fact that student perceptions were collected across test forms that included all the test items; however, when calculating differences in performance based on the student perceptions, we used the equated scores, which were based on the forms after items were dropped. It is possible that this could have impacted our results on student perceptions. That said, given that so few students got the difficult items correct, this was unlikely to impact the group-level results.

Lastly, another limitation to the study was the overall sample sizes across test forms. Given our small sample size, we were unable to evaluate differential item functioning (DIF) as a way to evaluate fairness of items across subgroups. As a result, future research should evaluate DIF for gender, race, and other subgroups to evaluate whether subgroups of examinees have different probabilities of success on an item after being matched on ability (Clauser and Mazor 1998). If items exhibit DIF they should be further analyzed to see what might be contributing to the DIF, or depending on the magnitude of the DIF, those items should be removed from the assessment.

Conclusions

This study provided preliminary evidence of the psychometric quality of the *HEIghten* Quantitative Literacy assessment. This study also provided insight to some potential gaps in quantitative literacy performance at higher education institutions, and points to the need to further investigate students' quantitative literacy skills. For instance, we found that students overall performed quite low on this assessment, that there were significant differences in performance across college majors, and a lack of learning gain from freshman to senior year of college. These results suggest that institutions may need to shift their current general education curriculum to require more quantitative literacy courses for all college students. Future research will benefit from further investigating these issues by working closely with higher education institutions to learn about their current curriculum and about their student population.

HEIghten Quantitative Literacy may be one way for institutions to capture information about students' quantitative literacy skills and to identify areas of gaps within the institution. *HEIghten* has the advantage of providing a clear construct definition to institutions. Given that this is a standardized assessment, institutions can also benchmark and compare their performance to other institutions using this assessment. In this study we demonstrated preliminary evidence to support the psychometric quality of this assessment, which is critical

when investigating student performance on a particular area such as quantitative literacy.

References

Abeli, Jamal. 1997. *Dimensionality of NAEP Subscale Scores in Mathematics.* Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

ACT. 2012. *ACT CAAP Technical Handbook 2011-2012.* Iowa City, IA: CAAP Program Management.

———. 2013. *ACT® –SAT® Concordance: A tool for comparing scores*. Accessed May 1, 2016. http://www.act.org/aap/concordance/pdf/reference.pdf.

Adelman, Cliff, Peter Ewell, Paul Gaston, and Carol Geary Schneider. 2014. *The Degree Qualifications Profile: A Learning-Centered Framework for What College Graduates Should Know and Be Able to Do to Earn the Associate, Bachelor's or Master's Degree.* Indianapolis, IN*:* Lumina Foundation for Education.

American Association of Community Colleges. 2016. *Principles and Plans: A Voluntary Framework of Accountability (VFA) for Community Colleges*. Accessed May 1, 2016, http://vfa.aacc.nche.edu/Documents/PrinciplesandPlans-AVoluntaryFrameworkofAccountability.pdf.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *The Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Arum, Richard, and Josipa Roksa. 2014. *Aspiring Adults Adrift: Tentative Transitions of College Graduates*. Chicago, IL: University of Chicago Press. https://doi.org/10.7208/chicago/9780226197142.001.0001.

Association of American Colleges and Universities. 2013. *The LEAP Vision for Learning: Outcomes, Practices, Impact, and Employers' view*. Washington, DC.

Allen, Mary J., and Wendy M. Yen. 1979. *Introduction to Measurement Theory.* Belmont, CA: Wadsworth.

Baer, Justin D., Andrea L. Cook, and Stephane Baldi. 2006. *The Literacy of America's College Students.* Washington, DC: American Institutes for Research.

Benjamin, Roger, Marc Chun, and Chris Jackson. 2009. *The Collegiate Learning Assessment's Place in the New Assessment and Accountability Space.* New York: Council for Aid to Education.

Blaich, C., and K. Wise. 2011. *From Gathering to Using Assessment Results: Lessons from the Wabash National Study. NILOA Occasional Paper #*8. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Clauser, Brian E., and Kathleen M. Mazor. 1998. "Using Statistical Procedures to Identify Differentially Functioning Test Items." *Educational Measurement: Issues and Practice* 17(1): 31–44. Accessed May 24, 2017. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Earlbaum Associates.

College Board. 2014. *Test Characteristics of the SAT: Reliability, Difficulty Levels, Completion Rates.* 2014. Accessed September 10, 2015, https://secure-media.collegeboard.org/digitalServices/pdf/sat/sat-characteristics-reliability-difficulty-completion-rates-2014.pdf.

Council for Aid to Education. 2015. *CLA+ technical FAQs.* New York, NY: Council for Aid to Education.

Dumford, Amber D. and Louis M. Rocconi. 2015. "Development of the Quantitative Reasoning Items on the National Survey of Student Engagement." *Numeracy* 8(1): Article 5. Accessed May 24, 2017. https://doi.org/10.5038/1936-4660.8.1.5.

Educational Testing Service. 2010. *ETS Proficiency Profile User's Guide.* (Princeton, NJ: Educational Testing Service.

Educational Testing Service. n.d. *HEIghten*^{TM} *Quantitative Literacy Sample Items.* 2017. Accessed May 25.2017. https://www.ets.org/s/heighten/pdf/quantitative-literacy-sample-questions.pdf.

Ekmekci, Adem. 2013. *Mathematical Literacy Assessment Design: A Dimensionality Analysis of Programme for International Student Assessment (PISA) Mathematics Framework.* PhD diss., University of Texas at Austin.

Elrod, Susan. 2014. "Quantitative Reasoning: The Next 'Across the Curriculum' Movement." *Peer Review* 16(3). Accessed May 24, 2017. https://www.aacu.org/peerreview/2014/summer/elrod.

Gall, Meredith Damien, Walter R. Borg, and Joyce P. Gall. 2007. *Educational Research: An Introduction.* Boston, MA: Pearson Education.

Gaze, Eric C., Aaron Montgomery, Semra Kilic-Bahi, Deann Leoni, Linda Misener, and Corrine Taylor. 2014. "Towards Developing a Quantitative Literacy/Reasoning Assessment Instrument." *Numeracy* 7(2): Article 4. Accessed May 24, 2017. https://doi.org/10.5038/1936-4660.7.2.4.

Goodman, Madeline, et al. 2013. *Literacy, Numeracy, and Problem Solving in Technology-Rich Environments among US Adults: Results from the Program for the International Assessment of Adult Competencies 2012.* NCES 2014-008. Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Graduate Management Admission Council. 2017. *Validity, Reliability, & Fairness*. 2017. Accessed May 25, 2017, http://www.gmac.com/gmat-other-assessments/about-the-gmat-exam/validity-reliability-fairness.aspx.

Hart Research Associates. 2015. *Falling short? College Learning and Career Success.* Washington, DC: Association of American Colleges and Universities.

Hughes-Hallett, Deborah. 2003. "The Role of Mathematics Courses in the Development of Quantitative Literacy." In *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*, edited by Bernard L. Madison and Lynn Arthur Steen, 91–98. Princeton, NJ: National Council on Education and the Disciplines.

Hyde, Janet S., Elizabeth Fennema, and Susan J. Lamon. 1990. "Gender Differences in Mathematics Performance: A Meta-Analysis." *Psychological Bulletin* 107(2): 139–155. https://doi.org/10.1037/0033-2909.107.2.139.

Klein, Stephen, Ou Lydia Liu, James Sconing, Roger Bolus, Brent Bridgeman, Heather Kugelmass, Alexander Nemeth, Steven Robbins, and Jeffrey Steedle. 2009. *Test Validity Study (TVS) report.* Washington, DC: Association of Public Land-Grant Universities.

Kolen, Michael J. "Traditional Equating Methodology." *Educational Measurement: Issues and Practice* 7(4): 29–37. Accessed May 24, 2017. https://doi.org/10.1111/j.1745-3992.1988.tb00843.x.

Kuncel, Nathan R., Sarah A. Hezlett, and Deniz S. Ones. 2001. "A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations: Implications for Graduate Student Selection and Performance." *Psychological Bulletin* 127(1): 162–181.

Lindberg, Sara M., Janet Shibley Hyde, Jennifer L. Petersen, and Marcia C. Linn. 2010. "New Trends in Gender and Mathematics Performance: A Meta-Analysis." *Psychological Bulletin* 13(6): 1123–1135. https://doi.org/10.1037/a0021276.

Livingston, Samuel A., and Sooyeon Kim. 2009. "The Circle-Arc Method for Equating in Small Samples." *Journal of Educational Measurement* 46(3) (2009): 330–343. Accessed May 24, 2017. https://doi.org/10.1111/j.1745-3984.2009.00084.x.

Liu, Ou Lydia. 2011. "Outcomes Assessment in Higher Education: Challenges and Future Research in the Context of Voluntary System of

Accountability." *Educational Measurement: Issues and Practice* 30(3): 2–9. Accessed May 24, 2017. https://doi.org/10.1111/j.1745-3992.2011.00206.x.

———, Brent Bridgeman, and Rachel M. Adler. 2012. "Measuring Learning Outcomes Assessment in Higher Education: Motivation Matters." *Educational Researcher* 41(9): 352–362.

Liu, Ou Lydia, Liyang Mao, Lois Frankel, and Jun Xu. 2016. "Assessing Critical Thinking in Higher Education: The HEIghten™ Approach and Preliminary Validity Evidence." *Assessment & Evaluation in Higher Education* 41(5): 677–694. Accessed May 24, 2017. https://doi.org/10.1080/02602938.2016.1168358.

Liu, Ou Lydia, Joseph A. Rios, and Victor Borden. 2015. "The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment." *Educational Assessment* 20(2): 79–94. Accessed May 24, 2017. https://doi.org/10.1080/10627197.2015.1028618.

Liu, Ou Lydia, and Katrina Crotts Roohr. 2013. *Investigating 10-Year Trends of Learning Outcomes at Community Colleges*. ETS RR-13-34. Princeton, NJ: Educational Testing Service.

Mislevy, Robert J., Linda S. Steinberg, and Russell G. Almond. 2002. *Design and Analysis in Task-Based Language Assessment*. CSE-TR-579. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing; California University; Center for the study of Evaluation.

Mislevy, Robert J., Russell G. Almond, and Janice F. Lukas. 2003. *A Brief Introduction to Evidence-Centered Design*. ETS RR-03-16. Princeton, NJ: Educational Testing Service.

Nunley, Charlene Rae, Trudy Haffron Bers, and Terri Manning. 2011. *Learning Outcomes Assessment in Community Colleges.* Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Organisation for Economic Co-Operation and Development. 2012. *Literacy, Numeracy and Problem-Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills.* Paris, France: OECD.

Rhodes, Terrel L., ed. 2010. *Assessing Outcomes and Improving Achievement: Tips and Tools for Using Rubrics.* Washington, DC: Association of American Colleges and Universities.

Rios, Joseph A., Ou Lydia Liu, and Brent Bridgeman. 2014. "Identifying Low-Effort Examinees on Student Learning Outcomes Assessment: A Comparison of Two Approaches." *New Directions for Institutional Research* 2014(161): 69–82. Accessed May 24, 2017. https://doi.org/10.1002/ir.20068.

Rock, Donald A.1991. "Subscale dimensionality." Paper resented at the annual meeting of the Design and Analysis Committee of the National Assessment of Educational Progress, Washington, DC, November.

Roohr, Katrina Crotts, Edith Aurora Graf, and Ou Lydia Liu. 2014. *Assessing Quantitative Literacy in Higher Education: An Overview of Existing Research and Assessments with Recommendations for Next-Generation Assessment.* ETS RR-14-22. Princeton, NJ: Educational Testing Service.

Shavelson, Richard J. 2008. "Reflections on Quantitative Reasoning: An assessment perspective." In *Calculation vs. Context: Quantitative Literacy and Its Implications for Teacher Education,* edited by Bernard L. Madison and Lynn Arthur Steen, 22–47. Washington, DC: Mathematical Association of America.

———. 2010. *Measuring College Learning Responsibly: Accountability in a New Era.* Stanford, CA: Stanford University Press).

Small, Don. 2003. "To Advance Quantitative Literacy, Improve College Algebra." In *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges,* edited by Bernard L. Madison and Lynn Arthur Steen, 252. Princeton, NJ: National Council on Education and the Disciplines.

Sons, Linda. ed. 1996. *Quantitative Reasoning for College Graduates: A Complement to the Standards.* Washington, DC: Mathematical Association of America.

Steen, Lynn A. 2001. "The Case for Quantitative Literacy." In *Mathematics and Democracy* edited by Lynn A. Steen, 1–22. Princeton, NJ: Woodrow Wilson National Fellowship Foundation, 2001.

Sundre, Donna L. 2008. *The Quantitative Reasoning Test, Version 9 (QR-9): Test Manual.* Harrisonburg, VA: The Center for Assessment & Research Studies.

U.S. Department of Education. 2014. *Have achievement gaps changed? The Nations Report Card*. Accessed February 10, 2017, https://nationsreportcard.gov/reading_math_2013/#/achievement-gaps.

Wise, Steven L. 2006. "An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test." *Applied Measurement in Education* 19(2): 95–114. Accessed May 24, 2017. https://doi.org/10.1207/s15324818ame1902_2.

———, and Christine E. DeMars. 2010. "Examinee Noneffort and the Validity of Program Assessment Results." *Educational Assessment* 15(1): 27–41. Accessed May 24, 2017. https://doi.org/10.1080/10627191003673216.

Wise, Steven L., and Xiaojing Kong. 2005. "Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests." *Applied Measurement in Education* 18(2): 163–183. Accessed May 24, 2017. https://doi.org/10.1207/s15324818ame1802_2.

Wise, Steven L., and Lingling Ma. 2012. "Setting response time thresholds for a CAT item pool: The normative threshold method." Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada. April 14-16.