

2011

# Quantitative Literacy at Michigan State University, 1: Development and Initial Evaluation of the Assessment

Alla Sikorskii

*Michigan State University*, sikorska@msu.edu

Vince Melfi

*Michigan State University*, melfi@stt.msu.edu

Dennis Gilliland

*Michigan State University*, gilliland@stt.msu.edu

Jennifer Kaplan

*Michigan State University*, jkaplan@uga.edu

Suzie Ahn

*Michigan State University*, suzie.ahn@hc.msu.edu

Follow this and additional works at: <http://scholarcommons.usf.edu/numeracy>



Part of the [Mathematics Commons](#), and the [Science and Mathematics Education Commons](#)

## Recommended Citation

Sikorskii, Alla; Melfi, Vince; Gilliland, Dennis; Kaplan, Jennifer; and Ahn, Suzie (2011) "Quantitative Literacy at Michigan State University, 1: Development and Initial Evaluation of the Assessment," *Numeracy*: Vol. 4 : Iss. 2 , Article 5.

DOI: <http://dx.doi.org/10.5038/1936-4660.4.2.5>

Available at: <http://scholarcommons.usf.edu/numeracy/vol4/iss2/art5>

---

# Quantitative Literacy at Michigan State University, 1: Development and Initial Evaluation of the Assessment

## Abstract

Development, psychometric testing, and the results of the administration of a quantitative literacy (QL) assessment to undergraduate students are described. Three forms were developed covering a wide range of skills, contexts, and quantitative information presentation formats. Following item generation and revision based on preliminary testing and cognitive interviewing, a total of 3,701 consented undergraduate students at Michigan State University completed one of the three forms. Two of the forms contained 14 multiple-choice items, and one form contained 17 multiple-choice items. All forms were completed by students in less than 30 minutes. Evidence of validity and reliability were obtained for the three forms. Unidimensionality of the underlying construct was established using confirmatory factor analysis. Correlations with ACT and university mathematics placement test ranged from .41 to .67, and correlations with the Lipkus numeracy scale ranged from .40 to .45. Cronbach's alphas for the three forms were near or exceeded .70. Comparison of student QL performance according to demographic characteristics revealed gender differences, with males scoring higher than females. These gender differences persisted even after controlling for ACT composite scores. Race/ethnicity differences were significant in unadjusted analysis, but did not persist over and above ACT composite scores in the adjusted analyses. The three newly developed forms of QL assessment will need to be further tested in the future to determine if they capture the effects of interventions that aim to improve QL.

## Keywords

quantitative literacy, numeracy, measurement, undergraduate education

## Creative Commons License



This work is licensed under a [Creative Commons Attribution-Noncommercial 4.0 License](https://creativecommons.org/licenses/by-nc/4.0/)

## Cover Page Footnote

Alla Sikorskii is Assistant Professor of Statistics and Probability at Michigan State University. Her research interests include quantitative literacy, methodology and measurement in health research, and stochastic processes.

Vince Melfi is Associate Professor of Statistics and Probability and Mathematics Education, and Director of the Mathematics Education Graduate Program at Michigan State University. His current research interests include quantitative literacy and adaptive allocation in clinical trials.

Dennis Gilliland is Professor of Statistics and Probability at Michigan State University. He is a Fellow of the American Statistical Association with current research interests in quantitative literacy, statistical inference and strategies for repeated games.

Jennifer J. Kaplan is an Assistant Professor of Statistics and Probability and Science and Mathematics Education at Michigan State University. Her research interests include undergraduate statistics education and quantitative literacy and adult statistical literacy.

---

Suzie Ahn received her Master's degree in Applied Statistics from Michigan State University and is a research assistant at Michigan State University Department of Statistics and Probability and College of Nursing.

## Introduction

In 2005 a University-wide task force at Michigan State University issued a report on Quantitative Literacy (QL) at the institution. In addition to providing a working definition of QL, the report made clear the need for ongoing assessment of students' QL skills (Estry and Ferrini-Mundy, 2005). This paper describes the assessment that was designed as a result of that report.

The task force report was produced at a time when assessment of student learning and accountability of schools were common themes in calls for improving the education of citizens. At the K–12 level, this included legislation such as the federal No Child Left Behind Act of 2001 and missions of advocacy groups such as Achieve.<sup>1</sup> Calls for assessment and accountability have recently become more common at the post-secondary level. Examples of higher education assessment and accountability requirements and recommendations for over twenty states are available on the Web site of the State Higher Education Executive Officers.<sup>2</sup> The board of directors of the Association of American Colleges and Universities has issued several calls for assessments to support accountability (e.g., AAC&U, 2010).

Quantitative literacy commonly has been cited as a learning outcome to be assessed in calls for assessment and accountability in higher education. For example the AAC&U report lists QL as one of six “Intellectual and Practical Skills” to be assessed (AAC&U, 2010).<sup>3</sup> These and similar efforts would benefit greatly from the availability of a psychometrically sound, easy-to-administer, and easy-to-grade assessment of QL targeted for use with undergraduate students. This manuscript reports, first, the development of such an assessment and, second, surprising gender differences in the QL scores that we found in testing it. A companion paper (Gilliland et al., 2011) uses the assessment to explore the correlation of QL with financial literacy.

### ***Definitions of Quantitative Literacy***

Quantitative Literacy (QL), also known as numeracy or quantitative reasoning, has been given a variety of operational definitions in the literature (Grawe and

---

<sup>1</sup> “an independent, bipartisan, non-profit education reform organization based in Washington, D.C. that helps states raise academic standards and graduation requirements, improve assessments and strengthen accountability.” <http://achieve.org/about-achieve> (accessed May 28, 2011).

<sup>2</sup> [http://www.shceo.org/links/links\\_results.asp?regionID=53&issueID=2](http://www.shceo.org/links/links_results.asp?regionID=53&issueID=2) (accessed May 28, 2011).

<sup>3</sup> The six skills referred to in the AAC&U report contain the ten skills referred to by the AAC&U in discussion of its “value rubrics.” For example, “critical thinking” and “creative thinking” have been combined into one category, “critical and creative thinking.” In either scheme, QL is one of the six, or ten, intellectual and practical skills to be assessed. AAC&U also provides a value rubric for QL with six core competencies such as Interpretation, Calculation, and Communication, and a future project will be to map the QL assessment items reported here to the AAC&U rubric.

Rutz, 2009). It is important to recognize at the outset that although QL makes use of mathematics, statistics, and other computational skills, it, like literacy, is not only a skill but also a “habit of mind”(Steen, 2001, p. 5).

The term “numeracy” was introduced in a 1959 report of the United Kingdom (UK) Committee on Education (Lloyd, 1959). A more recent UK report on school mathematics emphasized two attributes of numeracy: the ability to understand quantitative information, and the ability to use mathematical skills in everyday life (Cockroft, 1982). The first attribute reflects conceptual knowledge, a key to which is the understanding of principles; this understanding leads to the ability to use these principles in different contexts. The second attribute is related to procedural knowledge, defined as the ability to execute action sequences to solve problems (Rittle-Johnson et al., 2001),

As summarized in a recent review by Reyna et al. (2009) and papers by Sundre and Thelk (2010) and Wilkins (2010), other definitions agree that numeracy is practice-driven and involves more than mathematics and statistics knowledge (e.g., Adelsward and Sachs, 1996; Gal, 1995; Hallett, 2003; Madison and Steen, 2003; Montori and Rothman, 2005): the essential component of QL is an ability to choose and apply mathematical concepts in specific life and work situations that involve quantitative information. Steen (2004, p. 3) calls quantitative literacy a “uniquely modern blend of arithmetic with complex reasoning and (p. 4) describes the skills needed to be quantitatively literate as involving “sophisticated reasoning with elementary mathematics more often than elementary reasoning with sophisticated mathematics.” Thus QL assessments should differ from tests of mathematical knowledge and skills by including items that test the ability to reason and apply mathematical procedural skills in life situations.

### ***Existing Measures of QL and Related Constructs***

Existing measures of constructs related to QL include national and international surveys of educational achievement. The National Assessment of Education Progress (NAEP) in 2007 included the assessment of mathematical skills and their application to everyday life (Lee et al., 2007). While the 2003 Program for International Student Assessment (PISA) included questions relevant to QL, this and other surveys measure related constructs that are different from QL as defined above.<sup>4</sup> The U.S. Department of Education includes the assessment of QL as part of its literacy surveys by evaluating the ability to perform numerical operations and use numerical information in printed materials (Kirsch and Stein, 2002; Kutner et al., 2006).

---

<sup>4</sup> e.g., see <http://nces.ed.gov/surveys/pisa/> (accessed May 28, 2011).

Several measures of health-related QL have been developed (e.g., Schwartz et al. 1997; Lipkus et al., 2001; Estrada et al., 2004; Apter et al., 2006; Huizinga, Elasy et al., 2008; see Vacher and Chavez, 2009, for a review). Some of these measures are specific to particular health conditions (e.g., asthma numeracy skills, diabetes numeracy test) and are not suitable for QL assessment in the general student population of an educational institution. Other measures are disease-general such as the Schwartz-Woloshin questionnaire, which has three items that measure subjects' facility with probabilities, percents, and proportions (Schwartz et al., 1997). The Lipkus scale expands the Schwartz-Woloshin questionnaire to 11 items and covers risk and probability of a disease (Lipkus et al., 2001). Apart from asking about probability of a disease, the Lipkus scale has little health content, but is relevant to health and health decision-making. The items cover mainly probabilities, percents, and risks, and not the wider range of general mathematical and statistical knowledge and skills necessary for QL. Item response theory analysis of the Lipkus scale showed that most items had low difficulty. All but two had negative item location (difficulty) parameters which means that the Lipkus scale lacks items that can gauge QL levels that are above the mean (Schapira et al., 2009).

While the literature has multiple definitions of QL, and while constructs related to QL such as general literacy or mathematical knowledge and skills have received ample attention, there has been less empirical work published on assessment of QL among undergraduate students. To date there are a few examples of university assessments of QL. The University of Virginia, in order to satisfy requirements of the State Council of Higher Education in Virginia, assesses quantitative reasoning as one of its six core competencies (University of Virginia, 2007). Researchers at Carleton College (Grawe and Rutz, 2009), James Madison University (Sundre and Thelk, 2010) and Colby-Sawyer College (Steele and Kilic-Bahi, 2010) have developed quantitative inquiry, reasoning and knowledge assessments. Several other tests that assess quantitative reasoning, skills or statistical literacy are available (Carleton College, 2010). Our work complements and adds to these efforts and aims to develop and test a psychometrically sound QL assessment. Once this aim is achieved, the resulting assessment could be used with low respondent burden and high efficiency with respect to administration and scoring.

### ***QL at Michigan State University***

The development of the QL assessment reported here stemmed from a review of the QL requirements at Michigan State University (MSU). A task force with broad representation across the University was formed in 2004 to set MSU's goals for QL, to plan for assessment, and to make recommendations for changes to the curriculum. The task force defined QL succinctly as "the ability to formulate,

evaluate, and communicate conclusions and inferences from quantitative information” (Estry and Ferrini-Mundy, 2005). The task force elaborated on this definition:

Quantitative literacy employs analytical arguments and reasoning built upon fundamental concepts and skills of mathematics, statistics, and computing. Quantitatively literate MSU students will be more empowered members of a global society through their ability to represent and critique their world. (Estry and Ferrini-Mundy, 2005)

As seen from this and other definitions of QL, and as reported in empirical studies, QL is related to mathematics and statistics knowledge and level of education (Barnato et al., 2007; Galesic and Garcia-Retamero, 2010). As already noted, however, QL is distinct from these constructs.

The MSU task force report also described three stages of a college student’s development of QL: *prerequisite*, *foundational*, and *applied*. The prerequisite stage describes the QL expected of an entering student. The foundational stage describes the QL to be expected of all students prior to their admission to junior standing. The applied stage describes QL specific to the students’ discipline, which draws on the prerequisite and foundational QL stages. The assessments reported in this manuscript focus on the prerequisite stage of QL. In particular, items were mapped to middle and high school content expectations (standards) in the state of Michigan. As such, the instruments represent the QL ability expected of incoming freshmen at MSU.

Based on the QL Task Force, the working group on the assessment of quantitative literacy at MSU has identified five goals for QL at MSU:

1. Reading Graphical Displays: Read and interpret representations of quantitative information such as tables, charts and graphs.
2. Modeling Real-World Phenomena: Use multiple representations to model real-world phenomena.
3. Solving Real-World Problems: Solve real-world problems that require
  - a. application of geometric properties.
  - b. application of algebraic properties.
  - c. application of probability and statistics.
  - d. use of computer software.
  - e. critical thinking skills.
4. Justifying Conclusions: Explain and justify conclusions made with quantitative information including determining the appropriateness of the conclusion.
5. Critiquing Research Design: Critique the appropriateness of a research design relative to the research objectives.

## Materials and Methods

### *Item Selection*

The content of the QL assessment was informed by several considerations. First, the skills needed to respond to the items included:

- basic identification of quantitative information;
- proficiency in basic arithmetic and algebra and the ability to count, calculate and manipulate quantities;
- introductory proficiency in analytical arguments and reasoning and the ability to estimate and infer, possibly from several sources of quantitative information;
- knowledge of basic concepts of probability and statistics, critical appraisal of quantitative information such as probabilities of events, and strength of evidence obtained from various sources and studies with various designs.

These skills correspond to the four categories of health-related numeracy introduced by Golbeck et al. (2005): basic, computational, analytical and statistical health numeracy. This first consideration ensured the inclusion of items with a wide range of difficulty and coverage of the mathematical skills that exceeds that of existing assessments such as the Lipkus scale.

Second, the items were selected to reflect all five goals formulated by the MSU QL working group. Similar to the basic, computational, analytical and statistical categories of Golbeck et al. (2005), the MSU QL goals are not disjoint, but are inter-related. Therefore some of the items corresponded to more than one goal.

Third, the quantitative information presented in the items included numerical, graphical, statistical, and probabilistic information, which was presented using different formats: graphical displays (e.g., pie charts and graphs), probability presentations (e.g., natural frequencies and probabilities), and words and numbers. Recent research on the effects of formats on accuracy of understanding and communication (Gigerenzer et al., 2007; Reyna and Brainerd, 2007; Nelson et al., 2008; Galesic et al., 2009) suggests that some formats may be easier than others; thus the inclusion of items presenting information in different formats provided further partition of the levels of difficulty to capture differences in QL among students.

Fourth, a variety of contexts underlie the QL assessment items to reflect numerous life situations that involve quantitative information. These contexts included student life, financial, travel, sales and prices, home improvement, health and medical, and others. The items were set in these common contexts. It is worth noting, however, that these contexts are not completely authentic, in the sense that they have been adapted to remove ambiguity in wording and other problems that



are frequently encountered in authentic life situations. Future work will assess how well the scores on the QL assessments reported on here predict student performance on more authentic QL contexts, which cannot be captured by a multiple-choice assessment.

### ***Procedures of Assessment Development and Testing***

There were two phases to the instrument development and testing. Phase one consisted of the development and initial testing of the item pool. Phase two consisted of testing of the instruments that were assembled from individual items that were developed and tested during phase one.

The initial pool of over 50 candidate items was generated based on conceptual considerations and a review of the QL literature. Since the focus of the assessments was on the prerequisite stage of QL expected of entering freshmen (Estry and Ferrini-Mundy, 2005), each item was mapped to at least one of the Michigan Grade Level Content Expectations in Mathematics, which cover grades K–8,<sup>5</sup> or the Michigan High School Content Expectations in Mathematics, which cover grades 9–12.<sup>6</sup> Details of this mapping are lengthy, but are available from the authors by request.

Approval from the Institutional Review Board (IRB) was obtained for human subject testing of the items. The IRB applications were revised and amended as needed during the development process and approval was obtained prior to approaching any students. The items from the initial pool were administered in a series of assessments of students with various majors and seniority ranging from incoming freshmen to graduating seniors. These preliminary assessments allowed the investigators to evaluate item difficulty and possible sources of ambiguity in the wording of the items. For example, items with rates of either correct or incorrect responses exceeding 95% or items with negative correlations with total summed scores on other items were removed, as these items would not be useful in discriminating QL ability.

During this first stage of instrument development, the issues of instrument length and student motivation were discussed. Three different forms of a QL assessment were created: (1) a comprehensive *general form*, (2) a *basic form* that included primarily items assessing subjects' procedural fluency, and (3) an *advanced form* that included more items assessing student reasoning. The rationale for developing these three forms was to allow for a finer discrimination of the levels of QL with an instrument of reasonable length (30 minutes or less) that students would be willing to complete.

---

<sup>5</sup> [http://www.michigan.gov/mde/0,1607,7-140-28753\\_33232---,00.html](http://www.michigan.gov/mde/0,1607,7-140-28753_33232---,00.html) (retrieved May 18, 2011).

<sup>6</sup> [http://www.michigan.gov/mde/0,1607,7-140-38924\\_41644\\_42668---,00.html](http://www.michigan.gov/mde/0,1607,7-140-38924_41644_42668---,00.html) (retrieved May 18, 2011).

The second phase of the testing consisted of administering the forms to incoming freshmen during summer academic orientation sessions in 2009 and 2010. During this phase, the three above-mentioned QL forms were tested along with forms containing new candidate items. This testing reflected the ongoing process of generating an item bank of QL items that could be used in the future in similar assessments.

Without significant support from University administration, this project would not have been possible. The initial task force report (Estry and Ferrini-Mundy, 2005) was commissioned by the Office of the Provost. Financial support for development of the assessments was also provided by this office. Access to incoming students during summer academic orientation programs was arranged and supported by the Associate Provost and the Academic Orientation Office, and Academic Orientation staff assisted in delivering the assessments.

### ***Description, Scoring and Administration of the QL Assessments***

**Description.** The general QL form consists of 14 items; the basic QL form consists of 17 items; and the advanced QL form consists of 14 items. The three forms are in Appendix A (supplemental file).

**Scoring.** Items have multiple-choice format with the number of choices ranging from two to five. Each item was scored as either correct (1) or incorrect (0). The QL score was computed based on the sum of the individual item scores. The possible ranges of scores are from 0 (all items answered incorrectly) to 14 (all items answered correctly) for the general and advanced forms and from 0 to 17 for the basic QL form.

**Administration.** The investigators attended the academic orientation program sessions, which are mandatory for incoming freshmen, and handed out testing packets to the students. The enclosures consisted of a letter of explanation about the study, a consent form for the student to sign and to provide his or her name and student number so demographic and academic data could be obtained by the research team from University records,<sup>7</sup> the assessment itself, and an answer sheet on which the student recorded his or her answers. One of the investigators spoke to the students about the content of their packets and provided verbal explanation of the study according to a standardized script. Subsequently, the investigators provided the student numbers and responses to QL items to the appropriate University office and received the data set with demographic and

---

<sup>7</sup> After the first summer of assessment, investigators obtained IRB approval for bubbled-in consent in place of a signed consent form, and also for a waiver of consent for students who were 16 or 17 years of age at the time of assessment. These changes added to the sample size, helped assessment sessions to run more smoothly, and decreased paperwork significantly.

academic information added but with all personal identifiers (including name and student number) removed.

### ***Measures Obtained from the University Office***

To investigate the associations of QL scores with scores for related constructs, such as educational development, and to determine if differences in QL scores existed based on student characteristics such as gender, race, or declared major, a set of measures was obtained from the University office. *Socio-demographic variables* included gender, race, and ethnicity. *Academic variables* included current or intended major field of study, high school grade point average (GPA), standardized test scores (ACT and/or SAT), advanced placement (AP) calculus and statistics scores, and mathematics placement score. The intended majors were classified as Science, Technology, Engineering or Mathematics (STEM) versus non-STEM according to the National Center for Education Statistics classification of majors (Chen, 2009).

The mathematics placement test is administered by the Department of Mathematics at the time of entry into the University to assess students' mathematics knowledge and skills and to place students into mathematics courses of an appropriate level. Students have a maximum of two attempts to take the placement test without a proctor and one attempt to take a proctored test. The maximum score of all attempts was used in this study. This score ranges from 0 to 28 with higher scores indicating higher placement. Students with credit in AP calculus, or those with ACT mathematics score of 28 or higher or SAT mathematics score of 640 or higher were not required to take the mathematics placement test because these conditions were deemed by the Department of Mathematics to be sufficient for calculus placement.

### ***Lipkus Scale***

During some of the 2010 sessions, the Lipkus scale (Lipkus et al., 2001) was administered in addition to either the general, basic or advanced QL forms, to allow comparison of results on the two assessments. The Lipkus scale consists of 11 items, two in multiple-choice format and nine open-ended questions. All items are scored as correct or incorrect with possible scores ranging from 0 to 11. The items are listed in Lipkus, et al. (2001). A discussion of the Lipkus scale and its use in health numeracy studies is provided in Vacher and Chavez (2009).

### ***Sample***

The numbers of students who were approached, consented and completed each of the three forms are summarized in Table 1.

**Table 1**  
**Numbers of Students Who Were Approached, Consented, and Completed Each of the Three Forms of the QL Assessment\***

<i>Year</i>	<i>Total number approached</i>	<i>Total number consented</i>	<i>Completed general QL form</i>	<i>Completed basic QL form</i>	<i>Completed advanced QL form</i>
2009	6339	4221	257	445	364
2010	5338	5001	536	1582	517
<b>Total</b>	<b>11677</b>	<b>9222*</b>	<b>793</b>	<b>2027</b>	<b>881</b>

\*Several other QL and scientific reasoning assessments were also given, which accounts for the fact that 3701 (793+2027+881) of the 9222 consenting students completed one of the three QL forms reported on here

Of those approached in 2009, 1,757 were under 18 years of age and thus could not legally consent and 4,221 (92% of those 18 and older) signed the informed consent forms and completed various QL instruments. The consent rate in 2010 was 94%. In 2010 the IRB approved a waiver of consent for those 16 or over, so non-consenters were either under 16 or chose not to consent. This report is based on 3,701 incoming freshmen who completed one of the three QL forms in 2009-2010.<sup>8</sup>

### ***Approaches to Reliability and Validity Evaluations***

Content validity was initially established by the investigators who are disciplinary statisticians with considerable expertise in statistical education, quantitative literacy and decades of experience teaching K–12 and undergraduate statistics and mathematics courses. They developed and selected the items to be included in the assessment. As described above under item selection, items were selected to cover the areas of QL specified in the MSU QL Task Force definition and to meet the assessment of QL goals formulated by the QL working group.

One of the construct validity considerations is the dimensionality of the underlying construct of QL. It was hypothesized to be unidimensional, so that a single score indicating the number of items answered correctly is an adequate summary of the QL ability. Unidimensionality means that all items tap into the same dimension and reflect a one-dimensional construct of QL. To test the dimensionality, we conducted a confirmatory factor analysis (CFA), in which we tested if one underlying dimension was consistent with the data. Since item responses were scored as binary (correct or incorrect), CFA for categorical indicators was implemented in Mplus software (Muthen and Muthen, 1998-2007). Even though multiple indices of the goodness of fit exist, the root mean square

<sup>8</sup> Other QL assessments and an assessment of scientific reasoning were also given, which accounts for the fact that only 1066 students in 2009 and 2635 students in 2010 completed one of the three QL forms reported on here.

error of approximation (RMSEA) is often preferred since it is not inflated by the sample size (Kim, 2005). For this study a RMSEA of 0.05 or less was used as a goodness of fit criterion.

Item Response Theory (IRT) analyses were employed to assess the properties of items and their coverage of the underlying construct in relation to the potential range of the values of QL. IRT models evaluate whether a set of items can be used to measure, indirectly, a latent variable or construct (QL in this case), and also quantify items and estimates of students' underlying scores of QL in the same metric (Chang and Gehler, 2003). The IRT models use person and item parameters to explain the variation in persons' responses to different items. Person parameters in the IRT models are estimates of the latent variable (QL). Each item in the assessment has a set of numerical measures called item parameters. The probability of correct response to each item is expressed mathematically using the person's parameter and the parameters of the item. According to the number of parameters used to describe each item, IRT models are classified as one-, two-, and three-parameter. The first parameter, which is present in all three models, is the item difficulty, or item location parameter. "Easy" items have low item location parameters, and "difficult" items have high item location parameters. During the development of the assessment, the items should be chosen so that their range of difficulty covers the potential range of the latent variable (in this case, QL). A person whose IRT person parameter equals a value  $x$ , would be expected to answer correctly all items with item location parameters that are less than or equal to  $x$ . In practice, a person with high underlying value of QL may make a careless mistake and answer an easy item incorrectly. The IRT model we used is robust and can accommodate a small number of such unlikely events (Lord, 1980).

The second item parameter is the discrimination parameter, which reflects the fact that some items are more helpful in discriminating between those who have higher versus lower QL ability in a particular range. In addition, because the items have multiple-choice format, a guessing item parameter is included. Thus a three-parameter IRT model (named according to the number of parameters used for each item and also known as three-parameter logistic model) was fit using MULTILOG software, version 7.03.

The remaining analyses were performed in SAS 9.2. Since there is no "gold standard" measure of QL, the additional evidence of validity was limited to convergent validity (associations between measures of related yet different constructs). We evaluated the correlations between QL scores, mathematics placement test scores and ACT scores. In addition, 2010 data were used to evaluate the correlations between Lipkus scale scores and QL scores. Cronbach's alpha coefficient and between-item and item-total correlations were used to evaluate reliability. Additional evidence of validity comes from a study of the

relationship between QL and financial literacy reported in the companion paper (Gilliland et al., 2011).

Since differences in QL and related constructs have been reported according to race (Lemke et al., 2005; Ginde et al., 2008) and gender (Baker et al., 1999; Banks and Oldfield, 2007; Lusardi, 2008; Lusardi and Mitchell, 2009), we examined differences in QL scores by race and gender. In these analyses, the Native Americans were grouped with the “Other” race category because of low counts. To underscore the distinction between QL and the level of educational development and to separate the differences in QL from differences in educational development according to race and gender, we employed linear models with QL scores as dependent variables, and ACT mathematics or ACT composite scores, race and gender as covariates. Even though all students were incoming freshmen, their ACT scores had substantial variability as seen from the standard deviations reported in Table 3. ACT composite scores ranged from 13 to 36 in our sample; thus if differences in QL scores were found according to race or gender, the natural question was the attribution of these differences. In other words, are these differences due to the level of educational development? Controlling for ACT composite scores helped account for one of the possible sources of these differences. The linear models were implemented in the general linear model procedure (PROC GLM) in SAS 9.2. Least square (LS) means (or adjusted means) by race and gender were output from these models, and differences among them were tested.

## Results

### *Characteristics of the Sample*

The demographic characteristics of the 2009–2010 sample presented in Table 2 are consistent with summary data reported for all undergraduate students at the University. As seen from Table 2, no substantial demographic differences existed among those student groups that completed the three QL forms and the remaining group of students who completed other assessments. Thus, our sample is demographically representative of the incoming freshman population.

The academic characteristics and the means and standard deviations of QL scores are presented in Table 3. Again, no substantial academic differences are present among those student groups that completed the three QL forms and the remaining group of students who completed other assessments. These results suggest that our sample is academically representative of the incoming freshman population.

**Table 2**  
**Demographic Characteristics of Study Participants**

	<i>Entire 2009-2010 sample (N=9222)</i>	<i>General QL form sample (N=793)</i>	<i>Basic QL form sample (N=2027)</i>	<i>Advanced QL form sample (N=881)</i>
	<i>N (%)</i>	<i>N (%)</i>	<i>N (%)</i>	<i>N (%)</i>
<b>Race and Ethnicity</b>				
Caucasian/White non-Hispanic	7441 (80.7)	642 (81.0)	1688 (83.3)	680 (77.2)
Black non-Hispanic	650 (7.0)	57 (7.2)	104 (5.1)	76 (8.6)
Hispanic	356 (3.9)	30 (3.8)	88 (4.3)	41 (4.7)
Asian	412 (4.4)	27 (3.4)	82 (4.1)	47 (5.3)
Native American	53 (0.6)	5 (0.63)	8 (0.4)	4 (0.5)
Other	220 (2.4)	23 (2.9)	38 (1.9)	21 (2.4)
Refused or missing	90 (1.0)	9 (1.1)	19 (0.9)	12 (1.4)
<b>Gender</b>				
Female	4894 (53.1)	450 (56.8)	1088 (53.7)	502 (57.0)
Male	4246 (46.0)	334 (42.1)	922 (45.5)	374 (42.4)
Refused or missing	82 (0.89)	9 (1.1)	17 (0.8)	5 (0.6)

**Table 3**  
**Academic Characteristics of Study Participants: Means and Standard Deviations**

	<i>Entire 2009-2010 sample (N=9222)</i>	<i>General QL form sample (N=793)</i>	<i>Basic QL form sample (N=2027)</i>	<i>Advanced QL form sample (N=881)</i>
	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>	<i>Mean (SD)</i>
High school GPA	3.62 (0.34)	3.57 (0.34)	3.63 (0.33)	3.59 (0.36)
ACT English score	25.36 (4.48)	25.01 (4.47)	25.65 (4.47)	24.77 (4.58)
ACT Reading score	25.78 (4.86)	25.35 (4.93)	26.17 (4.85)	25.41 (5.03)
ACT Science score	24.99 (3.80)	24.75 (3.84)	25.21 (3.91)	24.54 (3.78)
ACT Mathematics score	25.18 (4.36)	24.70 (4.42)	25.38 (4.36)	24.77 (4.51)
ACT Writing score	24.45 (3.73)	24.20 (3.66)	24.80 (3.63)	23.96 (3.86)
ACT Composite score	25.49 (3.62)	25.11 (3.65)	25.76 (3.63)	25.03 (3.70)
Math. placement score	14.37 (5.78)	13.82 (5.74)	14.33 (5.76)	13.95 (5.77)
QL General	–	6.47 (2.85)	–	–
QL Basic	–	–	10.18 (3.24)	–
QL Advanced	–	–	–	6.48 (2.95)

SD = Standard deviation

## Findings Regarding Reliability and Validity

The internal consistency reliability of the QL forms was acceptable as reflected by Cronbach's alphas near .70. Cronbach's alphas for 2009, 2010 and combined data are presented in Table 4.

**Table 4**  
**Summary of the Internal Consistency Reliability (Cronbach's alphas)**

<i>Year</i>	<i>General QL form</i>	<i>Basic QL form</i>	<i>Advanced QL form</i>
2009	.70	.74	.72
2010	.69	.71	.65
2009-2010 combined	.69	.72	.68

The somewhat lower alphas for the general and advanced forms may be due to the wider range of knowledge and skills covered by these forms. Since the assessment was developed to test group differences and not individuals, the values of alpha are adequate (Wallace et al., 2009) and similar to the ones reported for other assessments (Steele and Kilic-Bahi, 2010). Item-total correlations varied across items because of varying item difficulty, and most were between .2 and .4.

The one-factor model fit was acceptable with RMSEA <.01 for the general form, RMSEA=.037 for the basic form, and RMSEA=.034 for the advanced form. Thus, the unidimensionality of the underlying construct was supported. Further evidence of validity was obtained from the correlations of the QL scores with Lipkus scores, mathematics placement scores and ACT scores (see Table 5). All

**Table 5.**  
**Correlations of QL Scores with ACT Scores and Mathematics Placement scores**

	<i>General QL score</i>	<i>Basic QL score</i>	<i>Advanced QL score</i>
ACT English score	.45 ( <i>n</i> =790)	.47 ( <i>n</i> =1966)	.47 ( <i>n</i> =833)
ACT Reading score	.43 ( <i>n</i> =760)	.41 ( <i>n</i> =1966)	.44 ( <i>n</i> =833)
ACT Science score	.61 ( <i>n</i> =760)	.59 ( <i>n</i> =1966)	.59 ( <i>n</i> =833)
ACT Mathematics score	.67 ( <i>n</i> =760)	.66 ( <i>n</i> =1966)	.65 ( <i>n</i> =833)
ACT Writing score	.41 ( <i>n</i> =653)	.43 ( <i>n</i> =1762)	.44 ( <i>n</i> =744)
ACT Composite score	.65 ( <i>n</i> =760)	.64 ( <i>n</i> =1966)	.65 ( <i>n</i> =833)
Mathematics placement score	.51 ( <i>n</i> =697)	.51 ( <i>n</i> =1776)	.53 ( <i>n</i> =787)
Lipkus score (subset of 2010 only)	.44 ( <i>n</i> =172)	.45 ( <i>n</i> =180)	.40 ( <i>n</i> =154)

*n* = number of students for whom the correlation was assessed

correlation coefficients were statistically different from zero (all *p*-values<.01). The magnitude of the correlations suggests that the construct of QL is related to knowledge in other subjects (e.g. mathematics and statistics), yet is distinct from these related constructs. It should be noted that the correlations with ACT



English, reading and writing were moderate and lower than correlations with ACT mathematics and ACT composite scores. The latter correlations ranged from .64 to .67 suggesting a strong association between QL scores and ACT mathematics and composite scores. An advantage of QL assessments over ACT tests stem from the facts that QL assessments are brief and will be available to researchers and educators.

Correlations of QL scores with Lipkus scores, while significant and supportive of concurrent validity of QL assessments, were moderate and ranged from .40 to .45. For comparison, correlations between Lipkus scores and ACT mathematics and composite scores ranged from .48 to .50 for the groups of students who completed the three forms of QL. Correlations between Lipkus scores and mathematics placement scores were between .25 and .30 (data not in tables).

The QL scores discriminated those with STEM versus non-STEM declared majors (Table 6). For all three forms, differences between STEM versus non-STEM groups were significant (all  $p$ -values  $<.0001$ ).

**Table 6**  
**Differences in QL Scores between STEM and Non-STEM Majors**

<i>Major classification</i>	<i>General QL form Mean (SD), N</i>	<i>Basic QL form Mean (SD), N</i>	<i>Advanced QL form Mean (SD), N</i>
STEM	7.39 (2.91), N=290	11.10 (3.33), N=739	7.11. (3.07), N=314
Non-STEM	5.93 (2.66), N=493	9.65 (3.09), N=1269	6.11 (2.82), N=558

SD=standard deviation

To evaluate the coverage of the range of potential QL values by the assessment items, item parameters estimated using IRT were examined. For the general QL form, percent of correct responses varied from item to item from 5% to 79%. Item location (difficulty) parameters derived from the IRT model with normal metric ranged from  $-1.31$  to  $3.20$ . In this metric, most values of the underlying QL would be expected to be between  $1$  and  $2$ , and essentially all between  $-3$  and  $3$ . Therefore the coverage of the potential values of the latent variable was deemed appropriate. For the basic QL form, percent of correct responses varied from 26% to 88%, and item location parameters covered the range from  $-2.98$  to  $1.50$ . For the advanced QL form, percents of correct responses were between 19% and 79%, and item location parameters ranged from  $-1.27$  to  $1.54$ .

### **Gender and Race/Ethnicity Differences in QL Scores**

Table 7 provides a summary of QL scores by gender, race and ethnicity. In unadjusted analyses that did not control for the ACT scores, differences by both race and gender were found: females had lower mean QL scores than males and

African American and Hispanic students had lower means compared to Caucasian non-Hispanic students. In the regression models that included race, gender and composite ACT score as covariates, differences by gender persisted. The effect of race, however, was no longer significant over and above gender and ACT scores.

Table 7

QL Scores by Race, Ethnicity and Gender, Unadjusted and Adjusted for ACT Composite Score

	<i>General QL score</i> (possible range 0-14)		<i>Basic QL score</i> (possible range 0-17)		<i>Advanced QL score</i> (possible range 0-14)	
	Mean (SE)	Adjusted Mean (SE)	Mean (SE)	Adjusted Mean (SE)	Mean (SE)	Adjusted Mean (SE)
Race and Ethnicity						
Caucasian non-Hispanic	6.74 (0.11)	6.60 (0.08)	10.47 (0.08)	10.28 (0.06)	6.77 (0.11)	6.53 (0.09)
Black non-Hispanic	<b>3.61 (0.36)*</b>	6.54 (0.30)	<b>6.88 (0.31)*</b>	10.17 (0.25)	<b>4.41 (0.33)*</b>	6.93 (0.28)
Asian	6.37 (0.53)	6.35 (0.40)	10.58 (0.34)	10.46 (0.26)	7.06 (0.42)	6.67 (0.33)
Hispanic	5.93 (0.50)	7.53 (0.41)	<b>8.34 (0.33)*</b>	10.09 (0.27)	<b>4.68 (0.45)*</b>	6.83 (0.37)
Other or missing	6.73 (0.45)	7.15 (0.39)	9.92 (0.39)	10.03 (0.35)	6.57 (0.47)	6.08 (0.41)
Gender						
Female	<b>5.54 (0.12)*</b>	<b>5.99 (0.16)*</b>	<b>9.21 (0.09)*</b>	<b>9.45 (0.13)*</b>	<b>5.79 (0.13)*</b>	<b>6.02 (0.16)*</b>
Male	7.72 (0.14)	7.68 (0.18)	11.34 (0.10)	10.96 (0.13)	7.40 (0.15)	7.19 (0.17)

SE=standard error

\*  $p < .01$  for differences with Caucasian race group in race/ethnicity comparisons or with males in gender comparisons.

Gender comparisons are adjusted for race/ethnicity and ACT composite scores; race/ethnicity comparisons are adjusted for gender and ACT composite scores.

We explored how the results presented in Table 7 would change if the ACT mathematics score was included as a covariate instead of the ACT composite score. No appreciable differences in the results were seen except that race remained significant after controlling for the ACT mathematics scores in the analysis of general QL form scores. Significant gender differences in QL scores persisted in all analyses.

The differences according to gender also existed in the ACT composite, ACT mathematics, and mathematics placement scores. The differences according to gender in mathematics placement scores were not significant over and above the ACT mathematics scores (data not shown). As mentioned earlier, since students with high ACT scores were not required to take mathematics placement test, mathematics placement scores were truncated at the higher end, and for this reason we used the ACT scores as covariates in the analysis of QL scores.

## Discussion and Conclusions

One common criticism of the assessment and accountability movement in higher education is that it is directed by government and/or administrations of colleges and universities, with little meaningful involvement of faculty members at the institutions (Lederman, 2010). These and other obstacles to faculty involvement along with suggestions for improvement are given by Hutchings (2010). In contrast, the development of the QL instruments at MSU has involved

collaboration between administration and faculty members, with administration providing context, such as the (faculty-driven) task force report, financial support, and, crucially, access to students at the academic orientation sessions, with faculty providing the expertise and efforts to create and test the instruments. This has proved to be a productive arrangement and may provide a model for assessment work at other institutions.

The newly created QL assessments begin to fill the gap in available instruments for assessing QL in the student population. Our results confirm that QL ability can not be assumed based on the level of education or performance on mathematics tests or standardized exams. While correlations with ACT mathematics or ACT composite scores were moderate to strong, QL assessments are shorter and capture more than the level of educational development. Differences in gender found over and above ACT composite scores further underscore the distinction between the construct of QL and level of education as well as mathematics skills and knowledge. Thus, valid and reliable assessments to measure QL of students at various stages of their undergraduate studies such as the ones described in this report are needed. These assessments can help identify gaps in student knowledge and skills so that educators can design programs and courses that specifically target these gaps.

Additional testing of the three QL forms is needed to establish their responsiveness, i.e., the ability to capture differences among groups of students and changes in QL over time or in response to the programs and interventions that aim to improve QL, such as such as those reported in Steele and Kilic-Bahi (2008, 2010). Interventions are urgently needed as low levels of QL in the general population have been documented (Lloyd, 1959; Schwartz et al., 1997, 2005; Estrada et al, 1999, 2004; Lipkus et al., 2001; Paulos, 2001;). A recent comparison of numeracy based on probabilistic samples from Germany and the U.S. revealed about 40% of the U.S. population have low numeracy, and 44% of people in the low numeracy group had at least some college education (Galesic and Garcia-Retamero, 2010). Thus adequate numeracy can not be assumed based on high school or college attendance; QL needs to be measured, and addressing quantitative illiteracy is an urgent need.

When race, ethnicity and gender status are considered, subgroups with even lower levels of QL have been identified. It was previously reported in Apter et al. (2006) that minorities had significantly lower scores on the numeracy questionnaire specific to asthma, but in the study of Apter and colleagues, the differences according to race virtually disappeared after controlling for the level of education. Race and ethnic disparities in numeracy have been identified previously (Lemke et al., 2005; Lusardi, 2008; Ginde et al., 2008), but in our study, there were no additional race or ethnicity effects above and beyond what is carried by the ACT composite scores, and only a small additional race effect for

the general form when controlling for ACT math scores rather than ACT composite scores.

The literature on gender differences in literacy and numeracy reports varying findings. Men were found to score significantly higher than women on the Test of Functional Health Literacy in Adults (TOFHLA) (Baker et al., 1999), on numeracy in the 2003 Adult Literacy and Lifeskills (ALL) Survey (Lemke et al. 2005), in the study of numeracy among adults 50 years of age and older in relation to understanding pensions (Banks and Oldfield, 2007), and among people of various age groups in financial literacy studies (Lusardi, 2008; Lusardi and Mitchell, 2009). At the same time, no differences were found by gender in scores on some other numeracy measures (Weiss et al., 2005; Apter et al., 2006), and a meta-analysis of 242 studies of mathematics performance published between 1990 and 2007 revealed similar performance of males and females (Lindberg et al., 2010). Similar findings were reported in meta-analysis of international studies of mathematics achievement, attitudes and affect; however the variability was high (Else-Quest et al., 2010). In our study differences by gender persisted after controlling for ACT scores, further supporting the conclusion of QL being distinct from educational development and raising a question as to the potential sources for the differences. Several papers attempted to explain gender differences in mathematics (e.g., Ernest, 1976; Halpern et al., 2007), and the sources of the differences in QL scores by gender will need to be further investigated in future work.

This study had several limitations. First, it was conducted at one large public university, and the results may not be generalizable to the entire population of undergraduate students. More work can be performed in the future to obtain additional evidence of validity by administering the assessment at other institutions. This paper is the first step in broader validation as it informs the scientific community of the instrument development and assessment work conducted to date. Second, while our sample was representative of the population of MSU students, it had somewhat low proportion of minority students. Third, this work does not address the responsiveness of the newly developed assessments. These are directions for future work that will be informed by the present study.

Finally we note that the need for sound assessments of quantitative literacy is not limited to educational institutions. A working paper by Gerardi and colleagues finds an association between low levels of QL and delinquency and default on subprime mortgages, based on a set of questions on quantitative literacy (Gerardi et al., 2010). Recent research demonstrated the impact of financial literacy including numeracy on economic outcomes for individuals (Banks and Oldfield, 2007; van Rooij et al., 2007; Lusardi, 2008; Lusardi and Mitchell, 2009) as well as health outcomes (Estrada, et al. 2004; Montori and Rothman, 2005; Donelle et

al., 2008; Huizinga, Beech et al., 2008; Waldrop-Valverde et al., 2009). With numerous educational, financial and health outcomes associated with quantitative literacy, the QL assessments described in this paper have the potential for use in research and practice of undergraduate education and beyond. Understanding students' QL abilities and how colleges and universities can improve students' QL abilities will require many and varied assessments and assessment strategies. We hope the work reported here will prove helpful as faculty members at other institutions engage in assessment of the QL of their students.

## Acknowledgements

**Research support.** Realizing the Vision, Quantitative Literacy, 2008-2011. Funds provided by the Michigan State University Office of Provost. D. Gilliland and V. Melfi, PIs.

The authors thank Donna L. Sundre, the anonymous referees, and the editor for their many useful suggestions. The authors take responsibility for any deficiencies remaining in the paper.

## References

- AAC&U. See Association of American Colleges and Universities.
- Adelsward, V., and L. Sachs. 1996. The meaning of 6.8: Numeracy and normality in health information talks. *Social Science and Medicine*, 43, 1179–1187. [http://dx.doi.org/10.1016/0277-9536\(95\)00366-5](http://dx.doi.org/10.1016/0277-9536(95)00366-5)
- Apter, A. J., J. Cheng, D. Small, I. M. Bennett, C. Albert, D. G. Fein, M. George, and S. Van Horne. 2006. Asthma numeracy skill and health literacy. *Journal of Asthma*, 43(9), 705–710. <http://dx.doi.org/10.1080/02770900600925585>
- Association of American Colleges and Universities. 2010. *Our students' best work: A framework for accountability worthy of our mission*. <http://www.aacu.org/publications/pdfs/studentsbestreport.pdf> (accessed May 28, 2011)
- Baker, D. W., M. V. Williams, R. M. Parker, J. A. Gazmararian, and J. Nurss. 1999. Development of a brief test to measure functional health literacy. *Patient Education and Counseling*, 38(1), 33–42. [http://dx.doi.org/10.1016/S0738-3991\(98\)00116-5](http://dx.doi.org/10.1016/S0738-3991(98)00116-5)
- Banks, J., and Z. Oldfield. 2007. Understanding pensions: Cognitive function, numerical ability, and retirement saving. *Fiscal Studies*, 28(2), 143–147. <http://dx.doi.org/10.1111/j.1475-5890.2007.00052.x>
- Barnato, A. E., H. A. Llewellyn-Thomas, E. M. Peters, L. Siminoff, E. D. Collins, and M. J. Barry. 2007. Communication and decision making in cancer care: Setting research priorities for decision support/ Patients' decision aids.

- Medical Decision Making*, 27, 626–633.  
<http://dx.doi.org/10.1177/0272989X07306788>
- Carleton College (2010). Carleton Quantitative Inquiry, Reasoning, & Knowledge Initiative. Assessment Tools. Retrieved May 12, 2011, from [http://serc.carleton.edu/quirk/pkal\\_workshop10/assessment\\_tools.html](http://serc.carleton.edu/quirk/pkal_workshop10/assessment_tools.html)
- Chang, C., and S. Gehlert. 2003. The Washington Psychological Seizure Inventory (WPSI): Psychometric evaluation and future applications. *Seizure*, 12, 261–267. [http://dx.doi.org/10.1016/S1059-1311\(02\)00275-3](http://dx.doi.org/10.1016/S1059-1311(02)00275-3)
- Chen, X. 2009. *Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education*. NCES Report 2009–161, <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009161> (accessed May 18, 2011).
- Cockroft, W. H. 1982. *Mathematics counts: Report of the Committee of Inquiry into the Teaching of Mathematics in Schools under the chairmanship of W.H. Cockroft*. London, England: Her Majesty's Stationery Office.
- Donelle, L., J. Arocha, and L. Hoffman-Goetz. 2008. Health literacy and numeracy: Key factors in cancer risk comprehension *Chronic Dis Can.*, 29(1), 1–8.
- Else-Quest, N. M., J. S. Hyde, J. S., and M. C. Linn, M. C. 2010. Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103–127. <http://dx.doi.org/10.1037/a0018053>
- Ernest, J. 1976. Mathematics and sex. *The American Mathematical Monthly*, 83(8), 595–614. <http://dx.doi.org/10.2307/2319884>
- Estrada, C. A., V. Barnes, C. Collins, and J. C. Byrd. 1999. Health literacy and numeracy. *JAMA*, 282(6), 523–529. <http://dx.doi.org/10.1001/jama.282.6.527>
- Estrada, C. A., M. Martin-Hryniewicz, B. T. Peek, C. Collins, and J. C. Byrd. 2004. Literacy and numeracy skills and anticoagulation control. *Am J Med Sci*, 328(2), 88–93.
- Estry, D. W., and J. Ferrini-Mundy. 2005. Quantitative Literacy Task Force Final Report and Recommendations. Michigan State University.
- Gal, I. 1995. Big picture: What does "numeracy" mean? *GED Items* 12(4/5), from <http://mathforum.org/teachers/adult.ed/articles/gal.html> (accessed May 28, 2011).
- Galesic, M., and R. Garcia-Retamero. 2010. Statistical numeracy for health. *Archives of Internal Medicine*, 170(5), 462–469. <http://dx.doi.org/10.1001/archinternmed.2009.481>
- Galesic, M., G. Gigerenzer, and N. Straubinger. 2009. Natural frequencies help older adults and people with low numeracy to evaluate medical screening

- tests. *Med Decis Making*.29(3), 368–371.  
<http://dx.doi.org/10.1177/0272989X08329463>
- Gerardi, K., L. Goette, and S. Meier. 2010. *Financial literacy and subprime mortgage delinquency: Evidence from a survey matched to administrative data*. Federal Reserve Bank of Atlanta Working Paper 2010-10.  
<http://www.frbatlanta.org/documents/pubs/wp/wp1010.pdf> (accessed May 28, 2011).
- Gigerenzer, G., W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. 2007. Helping doctors and patients make sense of health statistics. *Psychological Science in The Public Interest*, 8(2), 53–96.
- Gilliland, D., V. Melfi, A. Sikorskii, E. Corcoran, and E. Melfi. 2011. Quantitative literacy at Michigan State University, 2: Connection to financial literacy. *Numeracy*, 4(2), Article 6. <http://dx.doi.org/10.5038/1936-4660.4.2.6>
- Ginde, A., S. Clarke, J. Goldstein, and C. J. Camargo. 2008. Demographic disparities in numeracy among emergency department patients: Evidence from two multicenter studies. *Patient Educ Couns.*, 72(2), 350–356.  
<http://dx.doi.org/10.1016/j.pec.2008.03.012>
- Golbeck, A. L., C. R. Ahlers-Schmidt, A. M. Paschal, and S. E. Dismuke. 2005. A definition and operational framework for health numeracy. *American Journal of Preventive Medicine*, 29(4), 375–376.  
<http://dx.doi.org/10.1016/j.amepre.2005.06.012>
- Grawe, N. D., and C. A. Rutz. 2009. Integration with writing programs: A strategy for quantitative reasoning program development. *Numeracy*, 2(2), Article 2. <http://dx.doi.org/10.5038/1936-4660.2.2.2>
- Halpern, D. F., C. P. Benbow, D. C. Geary, R. C. Gur, J. S. Hyde, and M. A. Gernsbacher. 2007. The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, 8(1), 1–51.
- Hughes-Hallett, D. 2003. The Role of mathematics courses in the development of quantitative literacy. In *Why numeracy matters for schools and colleges*, ed. B. L. Madison and L. A. Steen, 91–98. Princeton, New Jersey: National Council on Education and the Disciplines.
- Huizinga, M. M., B. M. Beech, K. L. Cavanaugh, T. A. Elasy, and R. L. Rothman. 2008. Low numeracy skills are associated with higher BMI. *Obesity*, 16(8), 1966–1968. <http://dx.doi.org/10.1038/oby.2008.294>
- Huizinga, M. M., T. A. Elasy, K. A. Wallston, K. L. Cavanaugh, D. Davis, R. P. Gregory, L. S. Fuchs, R. Malone, A. Cherrington, D. A. Dewalt, J. Buse, M. Pignone, and R. L. Rothman. 2008. Development and validation of the Diabetes Numeracy Test (DNT). *BMC Health Services Research*, 8(96).
- Hutchings P. 2010. Opening doors to faculty involvement in assessment. National Institute for Learning Outcomes Assessment, Occasional Paper 4 (April

- 2010).  
<http://www.learningoutcomeassessment.org/documents/PatHutchings.pdf>.  
 (accessed May 28, 2011)
- Kim, K. H. 2005. The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, 12(3), 368–390.  
[http://dx.doi.org/10.1207/s15328007sem1203\\_2](http://dx.doi.org/10.1207/s15328007sem1203_2)
- Kirsch, J. and M. Stein, M. 2002. *Adult literacy in America: A first look at the findings of the National Adult Literacy Survey*. NCES Report 1993–275. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Kutner, M., E. Greenberg, Y. Jin, and C. Paulsen. 2006. *The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy* NCES Report 2006–483. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Lederman, D. 2010. The faculty role in assessment. *Inside Higher Education*.  
<http://www.insidehighered.com/news/2010/05/28/assess>. (accessed May 28, 2011).
- Lee, J., W. Grigg, and G. Dion. 2007. *The nation's report card: Mathematics 2007* NCES Report 2007–494. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Lemke, M., D. Miller, J. Johnston, T. Krenzke, L. Alvarez-Rojas, D. Kastberg, and L. Jocelyn. 2005. *Highlights from the 2003 International Adult Literacy and Lifeskills Survey (ALL)* (revised). NCES Publication 2005–5117.  
<http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005117> (accessed May 28, 2011).
- Lindberg, S., J. S. Hyde, J. L. Petersen, and M. C. Linn. 2010. New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123–1135. <http://dx.doi.org/10.1037/a0021276>
- Lipkus, I. M., G. Samsa, and B. K. Rimer. 2001. General performance on a numeracy scale among highly educated samples. *Med Decis Making*, 21, 37–44. <http://dx.doi.org/10.1177/0272989X0102100105>
- Lloyd, G. 1959. *Ministry of Education, 15 to 18; A report of the Central Advisory Council for Education (England) (Vol. 1)*. London, England: Her Majesty's Stationery Office.
- Lord, F. 1980. *Application of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Lusardi, A. 2008. Financial literacy: An essential tool for informed consumer choice? (Working paper UCC08-11).  
[http://www.jchs.harvard.edu/publications/finance/understanding\\_consumer\\_credit/papers/ucc08-11\\_lusardi.pdf](http://www.jchs.harvard.edu/publications/finance/understanding_consumer_credit/papers/ucc08-11_lusardi.pdf) (accessed May 28, 2011).



- , and O. S. Mitchell. 2009. How ordinary consumers make complex economic decisions: Financial literacy and retirement savings. National Bureau of Economic Research Working Paper No 15350. <http://www.nber.org/papers/w15350.pdf> (accessed May 28, 2011).
- Madison, B. L., and L. A. Steen. 2003. *Quantitative literacy: Why numeracy matters for schools and colleges*. Princeton, New Jersey: National Council on Education and the Disciplines.
- Montori, V., and R. Rothman. 2005. Weakness in numbers: The challenge of numeracy in health care. *J Gen Intern Med.*, 20(11), 1071–1072. <http://dx.doi.org/10.1111/j.1525-1497.2005.051498.x>
- Muthen, L. K., and B. O. Muthen. 1998–2007. Mplus user's guide. Los Angeles: Muthen & Muthen.
- Nelson, W., V. F. Reyna, A. Fagerlin, I. Lipkus, and E. Peters. 2008. Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine*, 35, 261–274. <http://dx.doi.org/10.1007/s12160-008-9037-8>
- Paulos, J. 2001. *Innumeracy: Mathematical illiteracy and its consequences*. New York: Hill and Wang.
- Reyna, V. F., and C. J. Brainerd. 2007. The importance of mathematics in health and human judgment: numeracy, risk communication, and medical decision making. *Learning and Individual Differences*, 17, 147–159. <http://dx.doi.org/10.1016/j.lindif.2007.03.010>
- Reyna, V. F., W. L. Nelson, P. K. Han, and N. F. Dieckmann. 2009. How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, 135(6), 943–973. <http://dx.doi.org/10.1037/a0017327>
- Rittle-Johnson, B., R. S. Siegler, and M. Alibali. 2001. Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, 93(2), 346–362. <http://dx.doi.org/10.1037/0022-0663.93.2.346>
- Schapira, M. M., C. M. Walker, and S. K. Sedivy. 2009. Evaluating existing measures of measures of health numeracy using item response theory. *Patient Education and Counseling*, 75(3), 308–314. <http://dx.doi.org/10.1016/j.pec.2009.03.035>
- Schwartz, L. M., S. Woloshin, W. C. Black, and H. G. Welch. 1997. The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med*, 127(11), 966–972.
- Schwartz, L. M., S. Woloshin, and H. G. Welch. 2005. Can patients interpret health information? An assessment of the medical data interpretation test. *Med Decis Making*, 25, 290–300. <http://dx.doi.org/10.1177/0272989X05276860>

- Steele, B., and S. Kilic-Bahi. 2008. Quantitative literacy across curriculum: A case study. *Numeracy*, 1(2), Article 3. <http://dx.doi.org/10.5038/1936-4660.1.2.3>
- . 2010. Quantitative literacy: Does it work? Evaluation of student outcomes at Colby-Sawyer College. *Numeracy*, 3(2), Article 3. <http://dx.doi.org/10.5038/1936-4660.3.2.3>
- Steen, L. A. 2001. *Mathematics and democracy: The case for quantitative literacy*. Princeton, New Jersey National Council on Education and the Disciplines
- . 2004. *Achieving QL: An urgent challenge for higher education*. Washington, DC: The Mathematical Association of America.
- Sundre, D., and A. D. Thelk. 2010. Advancing assessment of quantitative and scientific reasoning. . *Numeracy*, 3(2), article 2. <http://dx.doi.org/10.5038/1936-4660.3.2.2>
- University of Virginia. 2007. Reports: Core Competency Assessment Overview, <http://www.web.virginia.edu/iaas/reports/subject/competencies/overview.htm> (accessed May 28, 2011).
- Vacher, H. L., and T. Chavez. 2009. Quantitative literacy on the Web of Science, 2: Mining the health numeracy literature for assessment items. *Numeracy*, 2(1), Article 5. <http://dx.doi.org/10.5038/1936-4660.2.1.5>
- van Rooij, M., A. Lusardi, and R. Alessie. 2007. Financial literacy and stock market participation. National Bureau of Economic Research Working Paper No. W13565. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1024979##](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1024979##) (accessed May 28, 2011)
- Waldrop-Valverde, D., D. L. Jones, D. Jayaweera, P. Gonzalez, J. Romero, and R. L. Ownby. 2009. Gender difference in medication management capacity in HIV infection: The role of health literacy and numeracy. *AIDS Behav.*, 13(1), 46–52. <http://dx.doi.org/10.1007/s10461-008-9425-x>
- Wallace, D., K. Rheinlander, S. Woloshin, and L. Schwartz. 2009. Quantitative literacy assessments: An introduction to testing tests. *Numeracy*, 2(2), Article 3. <http://dx.doi.org/10.5038/1936-4660.2.2.3>
- Weiss, B. D., M. Z. Mays, W. Martz, K. M. Castro, D. A. DeWalt, M. P. Pignone, J. Mockbee, and F. A. Hale. 2005. Quick assessment of literacy in primary care: The newest vital sign. *Ann Fam Med*, 3, 514–522. <http://dx.doi.org/10.1370/afm.405>
- Wilkins, J. L. M. 2010. Modeling quantitative literacy. *Educational and Psychological Measurement*, 70(2), 267–290. <http://dx.doi.org/10.1177/0013164409344506>