Numeracy
Advancing Education in Quantitative Literacy

Volume 7 | Issue 2 Article 4

2014

Towards Developing a Quantitative Literacy/ Reasoning Assessment Instrument

Eric C. Gaze
Bowdoin College, egaze@bowdoin.edu
Aaron Montgomery
Central Washington University
Semra Kilic-Bahi
Colby-Sawyer College
Deann Leoni
Edmonds Community College
Linda Misener
Southern Maine Community College
Corrine Taylor
Wellesley College

Towards Developing a Quantitative Literacy/Reasoning Assessment Instrument

Abstract
This article reports on the development and implementation of a non-proprietary assessment instrument for Quantitative Literacy/Reasoning. This instrument was based on prior work by Bowdoin College, Colby-Sawyer College, and Wellesley College and was piloted in 2012 and 2013. This article presents a discussion of its development as well as the results of the pilot implementation. This work was supported by a TUES Type 1 grant from the National Science Foundation.

Keywords
quantitative literacy, quantitative reasoning, assessment, numeracy, test, instrument

Cover Page Footnote
Eric Gaze directs the Quantitative Reasoning program at Bowdoin College and is a Lecturer in the Mathematics Department. He is the current President of the National Numeracy Network (NNN 2013 – 2015). Prior to coming to Bowdoin, he led the development of a Masters in Numeracy program for K-12 teachers at Alfred University as an Associate Professor of Mathematics and Education.
Aaron Montgomery is Professor and former Chair of the Mathematics Department at Central Washington University. He is the current chair of the SIGMAA-QL (Special Interest Group of the Mathematical Association of America on Quantitative Literacy). He has developed curricula for the Carnegie Foundation's Quantway project, and the Dana Center's National Math Pathway's project.
Semra Kilic-Bahi is an Associate Professor of mathematics at Colby-Sawyer College within its Department of Natural Sciences. She is a past-chair of the SIGMAA-QL. She was Principal Investigator for the NSF project "Quantitative Literacy Across the Curriculum in a Liberal Arts Setting" and its supplement "Sustainability and Quantitative Literacy."
Deann Leoni is an Instructor in the Mathematics Department of Edmonds Community College (WA). She was a co-PI for the NSF projects "Mathematics Across the Curriculum" and "Mathematics Across the Community College Curriculum." Within these projects, she has team-taught interdisciplinary courses combining math with chemistry, English, political science, art, and art history.
Linda Misener is Associate Professor at Southern Maine Community College where she has taught a wide variety of topics including Developmental Math, Problem Solving, Discrete Mathematics, Statistics, Algebra, Trigonometry, Precalculus, and Calculus.
Corrine Taylor, an economist, has served as Director of the Quantitative Reasoning Program at Wellesley College since 2001. She is a past-president of the National Numeracy Network. Professor Taylor has led workshops, given invited lectures, and served as a consultant at other colleges and universities in the US and abroad that are developing new QR initiatives.

Introduction

Quantitative Literacy/Reasoning (QLR) has been in the academic landscape for over two decades. For the latter half of this period, academic institutions across the U.S. have been shifting the focus of introductory/general education math courses toward QLR, emphasizing the quantitative tools that students will need for successful decision making in their personal, professional, and civic lives (Schield 2010, Gaze 2014).

While QLR courses and curricula are finding wide dissemination, assessment of QLR in terms of skills of individual students and effectiveness of curricula remains primarily a local activity. There have been publications such as *Achieving Quantitative Literacy* (Steen 2004), and the AACU QL VALUE[1] rubric; however, most current assessment efforts are localized to a single campus, a single course, or even a single classroom (Taylor 2009). This is due, in part, to difficulties involved with assessing QLR skills (Wiggins 2003, Steen 2004, Boersma and Klyve 2013). Even at locations where QLR tests are implemented for placement purposes, tests are not being used for end-result assessment (Schield 2010).

Bowdoin College, Colby-Sawyer College, and Wellesley College have existing instruments that provided a starting point for this project. These tests have questions that cover four conceptual areas: number sense, reading and interpreting graphs, basic probability and statistics, and reasoning. Bowdoin and Colby-Sawyer assessment items are in selected-response (multiple-choice) format; the Wellesley College assessment items have an open-ended format. None of these instruments, however, allow for easy comparison across institutions since the instruments have been administered only locally. For example, Colby-Sawyer College developed and administered its QLR test to freshmen and seniors to assess and evaluate the impact of an NSF-supported *QL Across the Curriculum* initiative. At the end of the four-year evaluation process, the lack of national data on student QLR abilities left the Colby-Sawyer community to wonder about the level of impact of the initiative (Steele and Kilic-Bahi 2010). This dilemma is not new, nor is it restricted to Colby-Sawyer. The Director of the National Science Foundation expressed it succinctly when she stated:

> "We do not really know if we are making progress [since]...we do not have genuine benchmarks for what constitutes quantitative literacy." (Rita Carwell quoted in Steen 2004, p. 57)

[1] http://www.aacu.org/value/rubrics/index_p.cfm?CFID=52597980&CFTOKEN=88075268 (all links in the footnotes were accessed July 4, 2013)

The sentiment was echoed in 2008 in a paper in the *American Mathematical Monthly* which again stressed that most of these internal assessment tools have no national norms to compare to and the actual construct of "quantitatively literate" remains undeveloped (Bookman et al. 2008).

Purpose, Goals, and the QLR Construct

The QLR project described here aims to develop a valid and reliable test of QLR skills. In particular, the goals set forth in this NSF-supported project[2] were to design a QLR instrument that:

1. is non-proprietary,
2. provides a baseline of national QLR-scores from a variety of educational environments,
3. is reliable, and
4. has content validity.

We will start with the content validity piece, because the analysis of results to follow is meaningful only in the context of the construct we are attempting to measure. If the QLR construct is indeed "undeveloped" then how can we claim content validity? First and foremost is the experience of the QLRA project team. We have been teaching QLR courses, developing and authoring QLR curriculum materials, and assessing QLR skills for more than a decade. ECG is Director of the QR Center at Bowdoin College, past executive officer of the SIGMAA-QL[3] and current president of the NNN.[4] SKB was PI of the *QL Across the Curriculum* project at Colby-Sawyer and is a former executive officer of the SIGMAA-QL. DL (Edmonds Community College) was co-PI of the *Math Across the Curriculum* and the *Math across the Community College Curriculum* NSF projects. LM is QLR assessment coordinator at Southern Maine Community College. AM (Central Washington University) is the current SIGMAA-QL chair. CT is Director of the QR Program at Wellesley College and past-president of the NNN.

In fall 2011, the QLRA project team developed a 23-question QLR assessment instrument, the QLRA, by synthesizing the existing tests from Bowdoin, Colby-Sawyer, and Wellesley. Thirteen of the 23 questions came from Bowdoin's test, five came from Colby-Sawyer's test, and five were adapted from Wellesley's test. In addition to these 23 content questions, five survey questions, chosen from Dartmouth College's Math Attitudes Survey[5] and the Subjective

[2] Collaborative Research. Award Number 1140562, PI Eric Gaze, Co-PI Linda Misener, DUE, 2/15/2012; Award Number 1140584, PI Semra Kilic-Bahi, DUE, 2/15/2012.
[3] Special Interest Group, Mathematics Association of America, in Quantitative Literacy http://sigmaa.maa.org/ql/
[4] National Numeracy Network http://serc.carleton.edu/nnn/index.html
[5] http://www.math.dartmouth.edu/~matc/Evaluation/index.html

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

Numeracy Scale (Fagerlin et al. 2007), were included in the instrument and intended to measure attitudes towards mathematics and beliefs about one's mathematical ability. The 2012 and 2013 instruments are available upon request to ECG.

In discussing the questions we chose and why they are "QLR," it is helpful to have a working definition of QLR. Although there are many such definitions in the QLR community, the following is what we mean by QLR:

> *the skill set necessary to process quantitative information and the capacity to critique, reflect upon, and apply quantitative information in making decisions*.

It is interesting to note that cognitive psychologists have a similar definition for numeracy:

> A well-established and highly studied construct, numeracy encompasses not just mathematical ability but also a disposition to engage quantitative information in a reflective and systematic way and use it to support valid inferences. (Kahan et al. 2013)

Cognitive psychologists have been able to show that numeracy, the ability to use and understand numbers, is an effective scale for predicting behavior and decision making. In addition, cognitive reflection tests (CRT), which reflect one's ability to think beyond the first answer that comes to mind, have been shown to measure a different construct from numeracy (Liberali et al. 2012). Cognitive psychologists' dimensions of numeracy include quantitative skills along with deeper reasoning related to proportionality, matching, and relative magnitude. The QLRA project team members are not experts in assessment or cognitive processes, but are experts in pedagogy related to teaching QLR and have created the QLRA questions from this educational perspective. It is interesting that in comparing the QLRA questions to the numeracy scales and CRT of the cognitive psychologists we noted that the QLRA seems to be a combination of the two. We offer this observation merely to orient the QLRA relative to existing assessment scales and tasks. Our observation suggests an intriguing hypothesis for further study, but this paper aims simply to report the results from piloting the QLRA over two years.

The development of the QLRA is tied most closely to the Bowdoin test (13 out of the 23 questions in 2012). ECG had been refining the Bowdoin 30-question test over a three-year period (2009-12) as part of a joint project with Bates College funded by the Teagle Foundation.[6] There were several fundamental insights obtained from that project:

- Replace procedural, algorithmic questions with more involved reasoning, critical thinking questions.

---

[6] http://www.teaglefoundation.org/About/Mission-and-Vision

- Ask students to interpret tables and charts rather than doing it for them.
- Focus on quantitative literacy, using numbers in meaningful sentences rather than just computation.
- Ask students to postulate possible explanations for statistics rather than traditional logic games.

Content validity of the QLRA is also related to analyses carried out at Bowdoin as detailed below in the Discussion section. The Bowdoin test is highly correlated with both cumulative GPA and math/science GPA, even when controlling for other factors using multivariate regression. The correlation indicates the Bowdoin test is more than just a "math skills" test, compared to the math SAT which is not as highly correlated with cumulative GPA. Moreover, ECG has conducted pre- and post-course testing using the Bowdoin test in his QR course and has consistently seen his class improve one standard deviation, indicating that intentional teaching can improve QLR abilities.[7] These results raise the natural question of why not just use the Bowdoin test? To begin, we wanted this to be a collaborative project involving multiple institutions from the higher education spectrum: two-year colleges, public universities, and private liberal arts colleges. Given the selective admission at Bowdoin College and the wide range of schools the QLRA is intended to serve, concerns about a floor effect required that the Bowdoin test be adjusted with input from two-year schools and non-selective four-year schools. In addition, we needed to shorten the test so that it could be administered in class; this was especially important for community colleges.

The other three goals of the QLRA project are to create a non-proprietary, reliable test that would establish a national baseline of QLR abilities across the nation's higher education spectrum. Non-proprietary was again important in order to get the most buy-in, especially from public two- and four-year institutions. Reliability was measured for internal consistency using Cronbach's alpha, and also measured by comparing results across the two years of the pilot project and across institution types. The national baseline of QLR abilities will provide institutions with a reference point for their assessment efforts.

Potential purposes for the QLRA include:

- Advising students for course selection or placement into courses.
- Assessing QLR skills per se
- Evaluating courses and QLR curriculum initiatives
- Assessing summative outcomes including graduating students' QLR abilities.

[7]. Faculty in other disciplines have not conducted such pre-post testing so no comparisons are available

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

Overarching all of these is the hope to advance the methodology for measuring the construct of QLR.

Method

In spring 2012, the Quantitative Literacy and Reasoning Assessment (QLRA) was administered at 10 different institutions including community colleges, small liberal arts colleges, and large public universities. Students were recruited in a variety of ways that differed across institutions thus not allowing for a true random sample experimental design. Even though extra credit was the most commonly employed incentive (35%), monetary incentives ($10, $15, and $20 gift cards) were also used. The QLRA was administered in two modalities, a computer-based format and a paper-and-pencil format. The computer-based format was administered through online course-management systems such as Moodle or Blackboard. The use of calculators was allowed in all testing sites, but calculators were not provided at all sites. Test takers were either given 50–60 minutes to complete the test or allowed unlimited time depending on local scheduling opportunities. In general, most students seemed to finish within 30 to 45 minutes.

It is important to point out the challenges faced in recruitment and the decision to allow for non-uniform testing procedures. Getting schools to participate was difficult as was incentivizing the student participants. Community colleges had trouble getting clearance for incentive money. Faculty had difficulty getting participants even with incentives, which resulted in the test being administered in math classes. The QLRA project team wanted to pilot the instrument in as many diverse institutions as possible, and so allowances were made for disparate testing procedures. What we lacked in rigorous experimental design we made up for in participation rates. This trade-off has paid off, as word has spread about the test. In 2013, the QLRA online test site was completed as part of our website,[8] and already over 25 new schools are using the online platform in spring 2014, including many pre/post assessments. The online platform in particular will guarantee a uniform testing procedure.

In the summer of 2012, the QLRA project team refined the QLRA based on the pilot administration. Three problems with low item-total correlations were eliminated. One question was moved earlier in the test to determine if test fatigue explained the low score on the item. Other questions went through some minor changes in wording or presentation based on an analysis of student responses. Also, survey questions were added to the QLRA to better identify the demographics of the test subjects allowing data to be disaggregated based on

[8] http://serc.carleton.edu/qlra/index.html

gender, race, ethnicity, previous math courses taken, and year in school. This paper examines the results of the administration of the QLRA in both 2012 and 2013. The revised QLRA was administered in spring of 2013 at 11 different institutions including community colleges, small liberal arts colleges, and large public universities.

Factor analyses and item analyses were conducted to assess the internal consistency (Cronbach's alpha) and inter-item reliability. Post-hoc statistical tests were also used to determine whether gender or institution type affected student scores. In addition, tests were performed to determine the effect of converting Wellesley's open-ended questions to the multiple-choice format.

Subjects

In 2012, data were collected from 1,659 students and 10 institutions, and in 2013 data were collected from 2,173 students and 11 institutions. For 2014 we have over 25 new schools signed on to use the test and share results. Even though a standardized test administration protocol helped to reduce the variability in test administration, the use of the protocol was optional which means that cross-school differences may be the result of differences in recruiting practices or testing environment and not in student ability. Not all institutions provided demographic data.

The dataset was disaggregated by school type, sex, and, for the 2013 administration, graduation year (Table 1). Note that the datasets for sex and expected graduation date are smaller than the entire dataset because not all participants provided demographic information. Graduation year rather than class year was used as two-year schools do not have comparable class years.

Table 1
Subjects by Institution Type, Sex, and Graduation Year

		2012 *N* (%)	2013 *N* (%)
Institution Type	2-Year	314 (18.9)	273 (12.6)
	Non-selective 4-year	334 (20.1)	811 (37.3)
	Selective 4-year	1,011 (60.9)	1,088 (50.1)
	Total	**1,659 (100)**	**2,172 (100)**
Sex	Male	529 (50)	732 (39.7)
	Female	524 (50)	1111 (60.3)
	Total	**1,053 (100)**	**1,843 (100)**
Graduation Year	2013		472 (25.0)
	2014		80 (4.2)
	2015		192 (10.2)
	2016		647 (34.3)
	2017		488 (25.8)
	2018		10 (0.5)
	Total		**1,889 (100)**

Conceptual Areas

Prior to administering the test, the 23 items on the 2012 test were divided into four subscales: Number Sense (NS), Visual Representation (VR), Probability and

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

Statistics (PS), and Reasoning (R). Analysis of these areas did not reveal anything worth pursuing in this project.

Results

Descriptive Statistics

Overall results for the two administrations are given in Table 2 with histograms presented in Figure 1. Scores are provided as percentages in order to ease comparison between the 2012 results (23 questions) and the 2013 results (20 questions). A Kolmogorov-Smirnov test for normality was conducted with the results presented in Table 3. The test in both cases required the rejection of the null hypothesis ($p < .001$) and so normality of the data cannot be assumed. Note that the *x*-axis in the histogram is number of questions correct, not percent correct, to emphasize the different nature of the two tests. You can see the impact of more non-selective school participants in 2013 pulling the mean down.

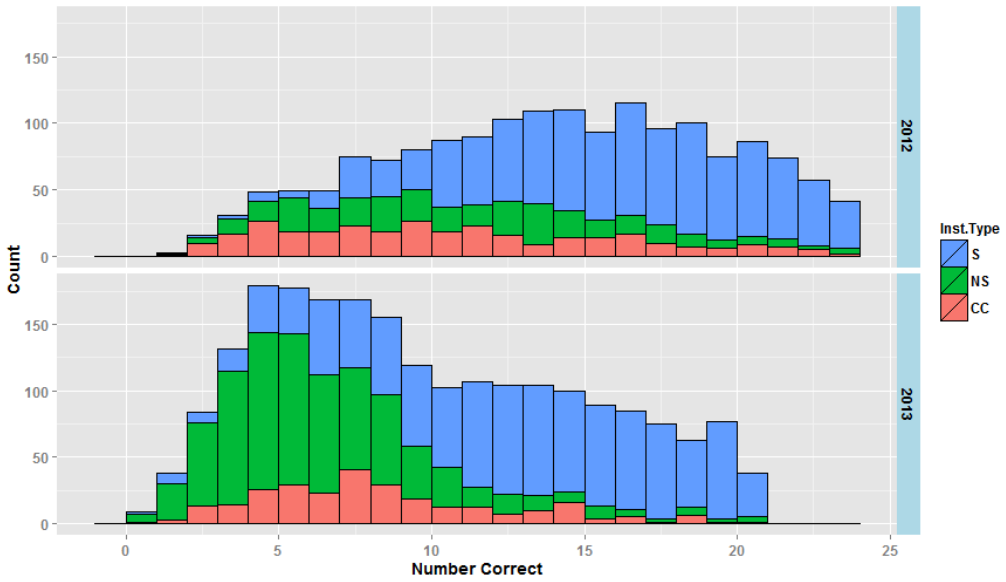

Figure 1. Histograms of QLRA scores by year and institution type. Key (institution type): S, four-year selective; NS, four-year non-selective; CC, community college. Note that the *x*-axis is number of questions correct rather than percentage correct because of the difference in the number of questions in the two tests. By inspection, the difference in the 2012 and 2013 distribution of results is associated with the increase in the proportion of the participants that were from NS institutions (see Tables 1 and 2).

Table 2

Descriptive Statistics, QLRA, 2012 and 2013

Year	*N*	Median (%)	Mean (%)	Std. Dev. (%)
2012	1659	60.9	58.4	23.2
2013	2172	40.0	46.1	25.1

Table 3

Results of Kolmogorov-Smirnov Test for Normality, QLRA, 2012 and 2013

Year	Statistic	df	Sig
2012	.072	1659	< .0005
2013	.112	2172	< .0005

Reliability

Overall reliability was tested using Cronbach's alpha, a statistic between 0 and 1 that increases as the inter-correlations of items increase. Thus it is a measure of the internal consistency or reliability that all items are measuring the same underlying construct. For the 2012 administration, Cronbach's alpha was 0.866, and, in 2013, it was 0.862.

Item-by-item analyses were performed on the 2012 and 2013 data, and results are in Tables A1 and A2 of the Appendix. The tables include the scale mean, variance, and Cronbach's alpha with the item deleted as well as the item-total correlations. Removal of any item in 2012 would decrease the value of Cronbach's alpha. Removal of any item in 2013 except Question 13 would decrease the value of Cronbach's alpha.

Item Difficulty and Distractors

Student success on each individual item was measured as the percentage of students who answered the problem correctly (Table 4). In 2012, values ranged from 26.8% to 86.5%, and, in 2013, they ranged from 24.8% to 73.2%.

Distractor analyses were performed on data from sites that returned item-specific information. This information was missing for approximately 12% of participants in 2012 and approximately 32% of participants in 2013; thus we had sufficient data for 1460 subjects in 2012 and 1478 subjects in 2013. The 2012 distractor analysis shows that the most-popular responses align with the correct answer in all but three questions (Questions 12, 17, and 22). The 2013 distractor analysis shows that the most-popular responses align with the correct answer except for the same three questions and question #1 (Questions 1, 12, 16 and 20). Questions where a distractor was the most-popular response are marked with an asterisk in Table 4, which shows the percentage of all subjects (not just those for

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

whom item information was available) who selected the correct answer and the percentage of all subjects who selected the most-popular distractor (for whom information was available).

Table 4
Item Difficulty and Distractor Analysis, QLRA, 2012 and 2013

Data for ALL Students				Data for Students with Item Information Given (*N*~1470)			
2012 Item	**2012 Mean %**	**2013 Number**	**2013 Mean %**	**2012 Correct %**	**2012 Mode Distractor %**	**2013 Correct %**	**2013 Mode Distractor %**
1	50.8(E)	1*	44.1(E)	49.6(E)	44.6(A)	36.3(E)	41.5(A)
2	44.2(B)	2	37.9(B)	42.8(B)	39.5(E)	44.8(B)	38.1(E)
3	80.8(A)	3	72.2(A)	81.6(A)	13.1(E)	79.3(A)	13.7(E)
4	51.2(E)	4	44.0(E)	50.4(E)	19.2(D)	50.2(E)	18.9(D)
5	85.0(D)	—	—	85.6(D)	6.4(E)	—	—
6	76.5(C)	6	65.9(C)	76.4(C)	12.3(D)	74.8(C)	13.2(D)
7	66.1(E)	7	53.2(E)	66.5(E)	12.5(C)	63.4(E)	14.3(C)
8	46.8(B)	8	37.3(B)	45.4(B)	21.2(E)	44.9(B)	21.2(E)
9	58.6(C)	9	47.5(C)	58.8(C)	14.8(A)	53.4(C)	16.5(A)
10	60.9(D)	10	49.1(D)	61.4(D)	33.7(C)	58.5(D)	35.5(C)
11	83.5(C)	11	73.2(C)	85.0(C)	7.2(D)	79.9(C)	10.0(D)
12*	30.9(E)	12*	27.3(E)	30.0(E)	40.9(C)	32.6(E)	41.8(C)
13	51.2(E)	13	42.9(E)	50.3(E)	20.5(C)	50.2(E)	16.1(B)
14	86.5(D)	—	—	87.3(D)	7.8(C)	—	—
15	66.8(B)	14	57.0(B)	66.7(B)	15.9(C)	66.9(B)	16.7(C)
16	73.2(B)	15	62.8(B)	73.8(B)	11.4(C)	74.2(B)	9.7(C)
17*	26.8(E)	16*	24.8(E)	26.4(E)	62.2(A)	32.6(E)	54.1(A)
18	56.2(A)	17	43.7(A)	56.0(A)	15.8(D)	54.0(A)	16.5(D)
19	58.2(D)	18	46.9(D)	59.2(D)	19.0(B)	55.6(D)	22.0(B)
20	55.6(B)	19	43.4(B)	55.3(B)	22.4(B)	53.7(B)	25.8(C)
21	64.0(D)	—	—	63.8(D)	19.9(E)	—	—
22*	36.9(D)	20*	30.1(D)	37.2(D)	45.7(E)	36.2(D)	45.5(E)
23	33.2(C)	5	27.7(C)	33.6(C)	19.9(A)	31.4(C)	30.2(D)

* Item for which the distractor is the most-popular response; otherwise, the correct answer was the most-popular response.

Inter-item correlations were computed for all items. Correlation values ranged from 0.04 to 0.41 for the 2012 test and from 0.075 to 0.384 for the 2013 test.

Exploratory Factor Analysis

The dimensionality of the 20 items on the 2013 QLRA was analyzed using maximum likelihood factor analysis. A scree test was conducted (Fig. 2) and indicated that the test was one-dimensional. Principal component analysis indicated that more than 28.132% of the variance could be accounted for from the first factor while the second factor accounted for only 5.873% of the variance. A similar factor analysis in for 2012 indicated the test was a unidimensional measure.

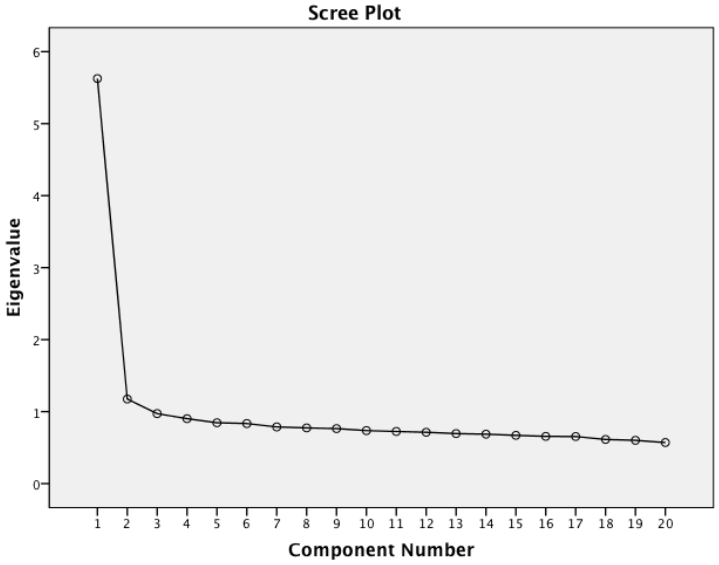

Figure 2. Scree plot from maximum likelihood factor analysis of the 20 items on the 2013 QLRA.

Using Item Response Theory (IRT), a further analysis was conducted on the 2012 data, and they fit a 3-PL model (difficulty, discrimination, and guessing parameters). Regarding difficulty, the point on the ability scale where the examinee has a 0.5 probability of correctly answering is designated as *b*, which is also known as the item difficulty parameter. That is, if an item has a *b*-parameter of 1.2, it means that an examinee with an ability estimate of 1.2 has a 50% chance of answering that question correctly. The range of *b*-parameters is on a continuum of –4 to 4, but it is most likely to see values in the range of –2 to 2.

Item discrimination is described by the *a*-parameter. Also known as the slope, the *a*-value indicates how discriminating the item is. Therefore, an item with a high *a*-parameter should be difficult for low-ability examinees and easy for high-ability examinees.

To take into account performance at the lower end of the ability spectrum, the *c*-parameter is added into the model. This is sometimes referred to as the "guessing parameter"; however, the *c*-value is usually slightly higher than the probability of randomly guessing the answer (because examinees usually rule out one or more options before choosing an answer). The values for this parameter should fall in the 0.15 to 0.25 range for questions with five response options.

For the QLRA, the average parameter values across all 23 items were $a = 1.027$; $b = -0.135$; and $c = 0.135$. Therefore, the items were, on average, discriminating and relatively easy.

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

Disaggregation

Data were disaggregated by institution type and gender for both years, and data were also disaggregated for expected graduation for 2013. Levene's statistic was computed with the following results: for 2012, Levene's statistic = 8.12 (1649); for 2013, Levene's statistic = 6.718 (1741). As a result, the null hypothesis of homogeneity of variance was rejected.

Disaggregated results by institution type are presented in Table 5. For the 2012 data, differences between institution types were found to be significant, $F = 1.88.494\ (1656),\ p < .001$. For the 2013 data, differences between institution types were found to be significant, $F = 481.863\ (2169),\ p < .0005$. Post-hoc testing used Tamhane's T2 statistics which indicated significant difference between each pair of institution types with $p < .0005$.

Table 5
Disaggregated Results by Institution Type, QLRA, 2012 and 2013

	2012			2013		
Institution Type	***N* (%)**	**Mean %**	**Std. Dev. %**	***N* (%)**	**Mean %**	**Std. Dev. %**
2-Year	314 (18.9)	44.7	23.4	273 (12.6)	39.3	20.2
Non-Selective 4-year	334 (20.1)	47.2	21.6	811 (37.3)	30.1	17.9
Selective 4-year	1011 (60.9)	66.4	20.0	1088 (50.1)	59.7	22.8
Total	**1659(100)**	**58.4**	**23.3**	**1659(100)**	**58.4**	**23.3**

Data were disaggregated by gender for both years and are presented in Table 6, although not all participants reported their gender. A Mann-Whitney U test was used to determine if differences in test scores were significant. For 2012, the test indicated significance, $z = -4.211,\ p < .001$; for 2013, the test indicated significance, $z = -7.693,\ p < .005$.

Table 6
Disaggregated Results by Gender, QLRA, 2012 and 2013

		2012			2013			
Gender	***N* (%)**	**Mean %**	**Median %**	**Std. Dev. %**	***N* (%)**	**Mean %**	**Median %**	**Std. Dev. %**
Male	529 (50.2)	62.9	65.2	22.7	732 (39.7)	49.6	50	23.9
Female	524 (49.8)	57.1	56.5	22.9	1111 (60.3)	41.2	35	23.8
Total	**1053 (100)**				**1843 (100)**			

Data were disaggregated by modality, Computer Based (CB) or Paper and Pencil (PP), and are presented in Table 7 and Figure 3. A Wilcoxon rank sum test with continuity correction revealed significant differences in modality scores for Non-Selective institutions for 2012 $(W = 40224,\ p < .0001)$, and Selective institutions for 2012 $(W = 51485,\ p < .0001)$ and 2013 $(W = 80496,\ p =$

0.005), but no significant difference for Non-Selective institutions for 2013 ($W = 40224,\ p = 0.8725$). The histogram in Figure 3 demonstrates resource issues faced by Community Colleges in not having information technology infrastructure to assist in computer-based test administration.

Table 7
Disaggregated Results by Modality, QLRA, 2012 and 2013

Institution Type	Modality	2012 *N* (%)	2012 Mean %	2012 Std. Dev. %	2013 *N* (%)	2013 Mean %	2013 Std. Dev. %
Non-Selective	Computer Based	65 (19.4)	36.3	16.7	695 (85.6)	30.3	18.6
	Paper and Pencil	270 (80.6)	49.7	21.9	117 (14.4)	28.8	13.3
Total		**335 (100)**			**812 (100)**		
Selective	Computer Based	768 (75.9)	62.7	19.3	880 (80.8)	59.6	23.6
	Paper and Pencil	244 (24.1)	78.1	17.4	209 (19.2)	64.8	21.9
Total		**1012(100)**			**1089(100)**		

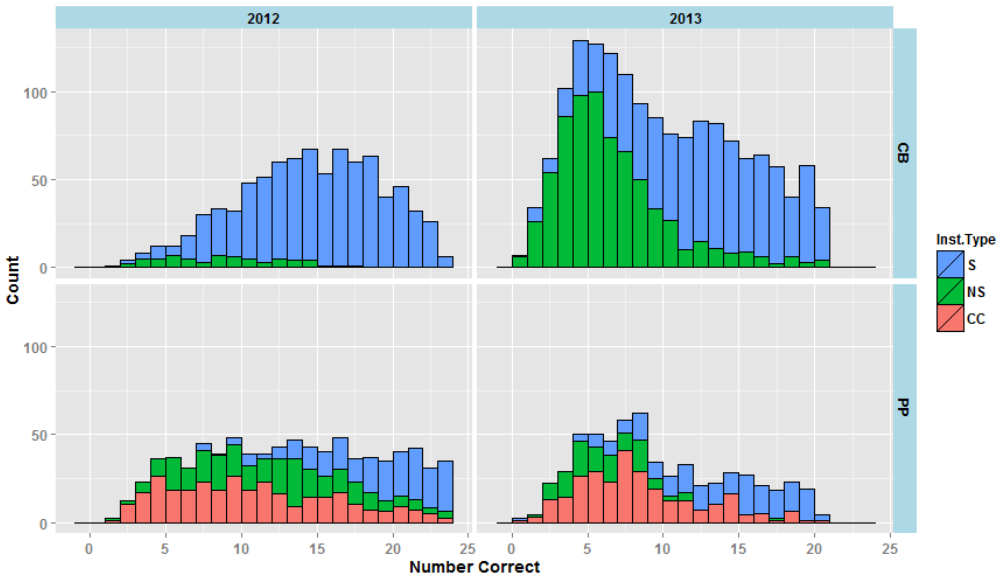

Figure 3. Histograms of disaggregated scores by year, institution type, and modality. Key (institution type): S, four-year selective; NS, four-year non-selective; CC, community college. Key (modality): CB, computer based; PP, paper and pencil.

Data were disaggregated by expected year of graduation for 2013 and are presented in Table 8, although not all participants reported their expected year of graduation. A Kruskal-Wallis test indicated there were significant differences between groups [chi-squared $(2,\ N = 1889) = 330.001,\ p < 0.005$]. Results of post-hoc testing using Tamhane's T2 are presented in Table A3 in the Appendix.

Table 8

Disaggregated Results by Expected Year of Graduation, QLRA, 2013

Year	*N* (%)	Mean %	Std. Dev. %
2013	472 (25)	46.8	26.2
2014	80 (4.2)	47.2	26.3
2015	192 (10.2)	54.0	26.9
2016	647 (34.3)	32.2	20.1
2017	488 (25.8)	56.1	20.2
2018	10 (0.5)	42.0	18.0
Total	**1889 (100)**		

Analyses of Attitude Survey Items

The following five survey items were included at the end of the 2013 instrument.

(1=Strongly Disagree, 5=Strongly Agree)

1. Numerical information is very useful in everyday life.
2. Numbers are not necessary for most situations.
3. Quantitative information is vital for accurate decisions.
4. Understanding numbers is as important in daily life as reading and writing.
5. It is a waste of time to learn information containing a lot of numbers.

Items #2 and #5 were reverse-scored so that high scores indicate high agreement that quantitative literacy and numeracy are important. The five attitude-survey items could function as a scale: Cronbach's alpha computed to .75, which is above the acceptable recommended value of .60 (IAR 2007)[9]. Means for each item and the total score are presented in Table 9, along with correlations to QLRA total score. For the overall group, the correlation between the two scores is $r = 0.37$. According to Cohen (1988), this value represents a medium-large effect size, indicating that attitude about QL does seem to relate to QLRA test score.

Table 9

Attitude-Survey Results, 2013

	N	Mean	SD	*r*
Total Survey	1076	18.65	3.51	0.37
Survey item 1	1106	3.75	1.02	0.36
Survey item 2	1104	3.43	0.97	0.09
Survey item 3	1106	3.66	0.91	0.37
Survey item 4	1104	3.81	1.04	0.24
Survey item 5	1103	3.97	0.99	0.27

[9] https://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php

Comparison with Wellesley's QLR Test

Three of the QLRA items were included as open-ended questions in Wellesley's QLR test. These items were given as open-ended questions to 608 Wellesley students and were given as multiple-choice items to 91 Wellesley students taking the QLRA. The open-ended questions on the Wellesley QLR test were graded dichotomously (either correct or incorrect) with no partial credit. The percentages of students who correctly answered the questions are included in Table 10.

Table 10
Comparison of QLRA Items to Wellesley QLR Open-Ended Test Items

	QLRA %		Wellesley %	
	2012	2013	2012	2013
Q1	63	81	77	81
Q8	66	65	37	37
Q10	76	87	89	90
N	91	83	608	610

Discussion

Substantial progress has been made on all four goals set forth in the Introduction. Institutions of various types can now assess their students using this non-proprietary instrument knowing that national comparisons are possible. Overall, the instrument is of sound quality. The content validity has been refined through the two pilot processes. The value of Cronbach's alpha (0.866 in 2012 and 0.862 in 2013) is an indication of very good reliability (DeVellis, 1991). This judgment is also supported by the values presented in Tables A1 and A2 where the item-total correlations exceed 0.20 for all items (IAR 2007). Thus reliability is high, item-total correlations and correlations among items are consistent with quality instrumentation, and item difficulties are within an appropriate range. Distractor analysis revealed that, for all but three items, the correct answer was chosen the most frequently.

Lack of normality for the sample may be an indication of the wide variety of participating institutions, as well as variations in recruiting methods. This lack of randomized stratified sampling leads to difficulty in interpretation of the statistics. In particular, Chi-squared tests indicate uneven distributions of gender across institution types [chi-squared (2, N=1843) = 69.806, $p < .0005$], expected years of graduation across institution types [chi-squared (10, N=1889) = 1319.155, $p < .0005$], and gender across expected year of graduation [chi-squared (5, N=1771) = 94.624, $p < .0005$].

Each of the three institution types has a mean that is statistically different than that of the other two institution types. This result is not surprising, as QLR

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

skills likely play a role in admission criteria for the different types of institutions. The result that the test indicates a difference between genders is a little surprising (although well documented), but of the three hypotheses (males have better QLR skills, the test is gender biased, or the sample is biased), we feel that sampling bias is the most likely explanation. More males came from selective schools and would graduate sooner than females. It may well be that this uneven distribution across schools and years of graduation accounts for the gender difference. The difference in scores due to modality (computer based versus paper and pencil) is intriguing and will need to be explored further. Further investigation is especially important as more schools utilize the online test administration site.

The comparison between the QLRA items and the Wellesley College open-ended version (Table 10) provides evidence that Q1 and Q10 are slightly easier as open-ended questions because there are no distractors to mislead students and the calculations are rather simple. Q8, a tax question, is far more complex, and our results seem to show that for more-complex problems – those where the table is harder to read and where there are more challenges in doing the calculations correctly – having a correct answer to choose from is easier than deriving that answer on one's own. Additionally, on the Wellesley College assessment, students are not permitted calculators, whereas they were allowed to use calculators on the QLRA. Also, for the QLRA, the Wellesley students had low stakes, whereas they were seriously invested in doing well on the Wellesley College assessment. The distractors, stakes, and use of calculators might all explain why students did better on the Wellesley College assessment for the simpler problems (although this discrepancy disappeared in the 2013 administration), but not as well on the harder, more calculation-intensive problem.

Comparison with Bowdoin's QLR Test

The significant overlap between the QLRA and Bowdoin's test provides further evidence of the soundness and validity of the QLRA. Bowdoin has conducted an extensive analysis of correlation between the score a student obtains on its test (the Q-score) and the student's academic performance. At Bowdoin, the Q-score is one of the best predictors of academic success. A student's Q-score is strongly correlated with the student's cumulative GPA (r = 0.39, N = 3,002 students from last 6 years), and MCSR GPA (r = 0.48) where MCSR represents Mathematical/Computational/Statistical Reasoning courses at Bowdoin. Also a student's Q-score is more strongly correlated (r = 0.48) with the student's first year cumulative GPA, again making the case for paying attention to this score in first-year advising and course selection.[10]

[10] Internal Document: *QR Academic Performance and Student Engagement in the MCSR Curriculum: Data Construction and Analysis*; D. Degraff, 2012

A multivariate analysis has also been conducted on Bowdoin's QLR test data. Multivariate regression allows one to control for multiple explanatory influences simultaneously. The first key result is that the models indicate that the Q-score is significantly predictive of Cumulative GPA and MCSR GPA even when controlling for a variety of other potential influences. This finding provides evidence that the associations indicated by the simpler bivariate correlations are likely to hold legitimate predictive power regarding future academic performance. Importantly, the models include Math and Verbal SAT/ACT scores among the explanatory variables. Thus, two entering students with identical Math and Verbal SAT scores, but different Q-scores, are predicted to have different GPAs. This finding suggests that there is additional information in the Q-score beyond that of the aptitude test scores and that the additional information would be of value for assessing future academic performance at Bowdoin. The fact that Q-score is at the high end of the range for both sets of coefficients (for cumulative and MCSR GPA) indicates the potential power of the Q-score to predict academic success across the curriculum, not just in MCSR courses.

In particular, see Table 11 for the coefficients from the multivariate regression model for both cumulative GPA and Math/Science (MCSR) GPA.[11] These coefficients indicate the predicted difference in GPA associated with a 10- percentage point increase in the respective aptitude test, with *all* other variables in the model held constant. Note that the verbal SAT score is the best predictor of cumulative GPA as might be expected, but Q-score is close behind. Math SAT is the best predictor for MCSR GPA, but again Q-score is close behind. Thus the QR test is measuring more than just quantitative skills but seems to be capturing deeper critical thinking skills. Table 12 shows how to interpret these coefficients for a 50-percentage point Q-score difference holding all other variables constant.

Table 11
Multivariate Regression Coefficients, Bowdoin QLR test (Q-Score)

	Math SAT	Q-Score	Verbal SAT
Cumulative GPA	0.0345	0.0603	0.0857
Math/Science (MCSR) GPA	0.1711	0.1599	0.0357

Table 12
Multivariate Model-Predicted Differences in GPA from a 50-percentage point change in Q-Score (Bowdoin QLR Test)

	Q-Score 30%	Q-Score 80%
Cumulative GPA	3.2	3.5
Math/Science (MCSR) GPA	2.7	3.5

[11] $r^2 = 0.30$ for Cum GPA and $r^2 = 0.36$ for MCSR GPA.

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

ECG has been using the Bowdoin QR exam for pre- and post-course testing in his QR class and has been able to improve students' Q-scores by approximately one standard deviation (Table 13).

Table 13
Improvement in Q-scores with Math 50, a QR Course(Bowdoin QRL Test)

	Pre-Course, Q-zscore	Post-Course, Q-zscore	Total Improvement
Math 50: QR Spring 2011			
Mean	-1.219	-0.253	0.966
StDev	0.905	0.913	
Math 50: QR Fall 2011			
Mean	-1.337	-0.210	1.127
StDev	0.670	0.913	
Math 50: QR Fall 2012			
Mean	-1.45	-0.230	0.916
StDev	0.694	0.607	

Refinements

In the summer of 2013, the QLRA project team continued to refine the test by adjusting the wording on some problems based on the spring 2013 results. The attitude-survey questions were moved to the beginning of the test, both to ensure that they would be answered and, hopefully, to inspire students to do their best on the exam. The three questions for which the most-frequent response was not the correct response were reworded in an attempt to reduce student confusion.

In the future, we will continue to fine-tune the wording and presentation of the tables and graphs that go with the questions as needed, and we will focus on extending the problem pool while retaining the reliability of the test. Now that the test is being mandated at some institutions, the need to recruit participants through incentives has been greatly reduced. This should allow for more standardized testing environments across schools which will improve the quality of the baseline data.

Conclusion

We believe that the QLRA is a valid measure of Quantitative Literacy/Reasoning given its construction by practicing experts in the field of QR, its internal consistency, item coding of questions, and associated correlation to Bowdoin's math/science and cumulative GPA. Reliability has been demonstrated by consistency over two years of pilot data and across multiple institutions with different student demographics. Intentional teaching in a QR course has been

shown to improve student performance using pre-post testing; indicating relevance of such curriculum to improving academic performance of students. Further research relating the QLRA to Numeracy tests and Cognitive Reflection Tests from the field of Cognitive Science is warranted. The QLRA is a simple, easy-to-use tool, providing powerful data for student advising and placement in courses.

References

Boersma, S., and D. Klyve. 2013. Measuring habits of mind: Toward a prompt-less instrment for assessing quantitative literacy. *Numeracy* 6 (1): Article 6. http://dx.doi.org/10.5038/1936-4660.6.1.6

Bookman, J., S. Ganter, and R. Morgan. 2008. Developing assessment methodologies for quantitative literacy: A formative study. *American Mathematical Monthly* 115 (10): 911-929.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale NJ: Lawrence Erlbaum Associates.

DeVellis, R. F. 1991. *Scale development: Theory and applications*. Newbury Park CA: Sage Publications.

Fagerlin, A., B.J. Zigmund-Fisher, P.A. Ubel, A. Jankovic, H.A. Derry, and D.M. Smith. 2007. Measuring numeracy without a math test: Development of the Subjective Numeracy Scale (SNS). *Medical Decision Making* 27 (5): 672-680. http://dx.doi.org/10.1177/0272989X07304449

Gaze, E. 2014. Teaching quantitative reasoning: A better context for algebra. *Numeracy* 7 (1): Article 1. http://dx.doi.org/10.5038/1936-4660.7.1.1

IAR. See Instructional Assessment Resources. 2007.

Instructional Assessment Resources. 2007. Assess students: Item analysis. The University of Texas at Austin. https://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php Last updated Sept 21, 2011 (accessed June 5, 2014).

Kahan, D., E. Dawson, E. Peters, and P. Slovic. 2013. Motivated numeracy and enlightened self-government. The Cognition Project Working Paper No. 116. Yale Law School.

Liberali, J., V. Reyna, S. Furlan, L. Stein, and S. Pardo. 2012. Individual differences in numeracy and cognitive reflection with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making* 25: 361-381. http://dx.doi.org/10.1002/bdm.752

Schield, M. 2010. Assessing statistical literacy: Take CARE. In *Assessment methods in statistical education: An international perspective*, ed. P. Bidgood, N. Hunt, and Flaiva Jolliffe, 133-152. Hoboken NJ: John Wiley & Sons.

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4

Steele, B. and S. Kilic-Bahi. 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. 2004. *Achieving quantitative literacy: An urgent challenge for higher education.* Washington D.C.: The Mathematical Association of America.

Taylor, C. 2009. Assessing quantitative reasoning. *Numeracy* 2 (2): Article 1. http://dx.doi.org/10.5038/1936-4660.2.2.1

Appendix

Table A1

2012 Item-by-Item Analyses with Corrected Item-Total Correlations

2012 Number	2013 Number	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	1	12.93	26.233	.425	.861
2	2	13.00	26.066	.463	.860
3	3	12.63	26.994	.369	.863
4	4	12.93	26.265	.418	.861
5	—	12.59	27.160	.369	.863
6	6	12.67	26.518	.449	.860
7	7	12.78	25.830	.541	.857
8	8	12.97	26.191	.435	.861
9	9	12.85	26.469	.384	.863
10	10	12.83	25.930	.501	.859
11	11	12.60	27.161	.352	.863
12	12	13.13	26.162	.483	.859
13	13	12.93	26.389	.393	.862
14	—	12.57	27.374	.327	.864
15	14	12.77	26.558	.387	.862
16	15	12.71	26.249	.488	.859
17	16	13.17	26.287	.479	.859
18	17	12.88	25.627	.555	.857
19	18	12.86	26.059	.468	.860
20	19	12.88	25.424	.596	.855
21	—	12.80	26.456	.400	.862
22	20	13.07	26.311	.427	.861
23	5	13.11	27.072	.279	.866

Table A2

2013 Item-by-Item Analyses with Corrected Item-Total Correlations

2013 Number	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	8.78	22.76	0.47	0.855
2	8.84	22.91	0.45	0.856
3	8.50	23.33	0.40	0.858
4	8.78	22.97	0.43	0.857
5	8.94	23.62	0.33	0.860
6	8.56	22.98	0.45	0.856
7	8.69	22.46	0.54	0.852
8	8.85	22.96	0.44	0.856
9	8.74	23.02	0.41	0.857
10	8.73	22.50	0.52	0.853
11	8.49	23.29	0.41	0.857
12	8.95	23.33	0.40	0.858
13	8.88	23.97	0.23	0.864
14	8.65	22.75	0.48	0.855
15	8.59	22.58	0.53	0.853
16	8.97	22.92	0.52	0.854
17	8.78	22.34	0.57	0.851
18	8.75	22.72	0.48	0.855
19	8.78	22.38	0.56	0.851
20	8.92	23.13	0.43	0.856

Table A3

Tamhane’s T2 Significance between Expected Years of Graduation

	2013	2014	2015	2016	2017	2018
2013	1.000	.025*	<.0005*	<.0005*	.0005*	1.000
2014	—	1.000	.580	<.0005*	.067	1.000
2015	—	—	1.000	<.0005*	.997	.671
2016	—	—	—	1.000	<.0005*	.856
2017	—	—	—	—	1.000	.418
2018	—	—	—	—	—	1.000

DOI: http://dx.doi.org/10.5038/1936-4660.7.2.4