



Lessons Learned about Assessing Quantitative Literacy

MAA PREP Workshop:
Creating and Strengthening Interdisciplinary Programs in
Quantitative Literacy



June 14–18, 2005, Macalester College, St. Paul, Minnesota

Jack Bookman
Mathematics Department
Duke University
bookman@math.duke.edu



Assessing QL presents the same challenges of assessing other types of learning but QL presents some particular challenges for assessment:

- Assessment items must be set in a real world context.
- The problems and the contexts must be familiar to all the students but neither so routine that they require very little thought nor so non-routine that students cannot solve them.
- The assessment instruments must not take too much time to administer, must include a multitude of problems and situations, and must be designed to allow for reliable scoring.

The "lessons learned" have come from work on two evaluation projects:

- Macalester College's Quantitative Methods for Public Policy
 - the goal, that students can "demonstrate basic understanding of the mathematical language of change and comparison" will be measured using pre-and post-tests
- An NSF project where we examined some existing QL tests used at two liberal arts colleges
 - item analyses of the placement tests
 - pre-test and post-test results from the QL course






Making up good test items is harder than we thought

In the first attempt at making up questions for QM4PP assessment, we discovered a problem of false negatives – a student who may know the quantitative implications of a question may not address those issues in his or her answer.

Example – For the statement below, what additional thing(s) would you need to know to make a decision?



Because immigrants pay taxes, we should allow anyone into the US as an immigrant.

Making up good test items is harder than we thought

Too little guidance should be avoided as seen in the previous question.

Too much guidance should be avoided.

Making up good test items is harder than we thought

Students may surprise you with their answers

For example:

The government will allow people age 55 to prepay their retirement taxes --- the taxes that they need to pay on the retirement savings that they withdraw or on social security payments. For every \$1000 that someone would pay at age 65, the potential retiree can pre-pay, at age 55, at a discount: \$500. Does this make financial sense for the government to do this? Briefly say why.

Coming up with a good grading rubric was harder than we thought

We used the following general rubric for scoring:
 4: Top score. The three C's need to be there: Correct, Complete, Clear.
 3: Close but not quite. Generally missing one of the three C's earned a 3.
 2: Partial. Best possible for an incorrect answer, worst possible for a correct answer. Often a good answer with no support or a wrong answer with good but misguided support.
 1: Attempt was made.
 0: Blank or a variant of "I have no idea" with no actual try.

A more specific rubric was then formed for each of the fourteen questions.

In other more standard math tests of concepts or problem solving, this type of rubric has been reasonably effective with high inter-rater reliability. We found this harder to do with the more nuanced questions in our QL tests.

Reaching proficiency may take more than a one semester course

For each of the two schools in the NSF study, we examined the pre and post test results for those students who did not pass the proficiency test and were placed into a quantitative literacy course.

We found (at each of the institutions) statistically significant growth during the one semester QL course, even after scores were adjusted to consider regression toward the mean.

However, in spite of this growth, the average post-test score was still below the competency level as defined for the pre-test. This is consistent with results reported elsewhere (at Virginia Commonwealth for example).

We believe that this should be expected since there is only so much growth that one can expect from one course. We believe that this gives some preliminary evidence for the assertion that in order to create a large positive effect concerning QL, colleges and universities must have a sustained effort that continues past a one course requirement.

There are some reasonably easy statistical tests that can be used to examine how well individual items serve the goal of the test.

In particular, if the goal of the test is to be a proficiency test it is possible to examine how well a question discriminates at the passing level by plotting the scores on each question vs. the total score on the test.

In order to examine the ability of the questions to discriminate at the critical score of 30, we plotted the scores on each question (adjusted for difficulty level, that is, the score on the question/maximum possible score for that question) vs. the total score on the test and fit a cubic equation to that scatterplot. The greater the slope at the critical score of 30, the better the question discriminates across students at that criterion. For example, compare the graphs in Figure 1 below for questions 4 and 5 and notice how much better question 4 discriminates at QR = 30.

One can look at the difficulty level for each of the questions. Note that 1, 5, 7, and 11 below are relatively easy questions and are therefore less likely to discriminate between stronger and weaker students.

	Mean	Std. Deviation
RPRE1	0.82833	0.29319
RPRE2	0.56	0.438591
RPRE3	0.3875	0.474044
RPRE4	0.595	0.38953
RPRE5	0.8525	0.26899
RPRE6	0.53667	0.370296
RPRE7	0.85	0.357967
RPRE8	0.45125	0.323588
RPRE9	0.72125	0.353711
RPRE10	0.4925	0.319891
RPRE11	0.8925	0.194958
RPRE12	0.61375	0.369875
RPRE13	0.575	0.395307
RPRE14	0.345	0.463193
RPRE15	0.69875	0.384448

We also computed the correlation coefficient for each question with the total score less the score on that question. Note that questions 5 and 7 below correlate rather poorly with the other questions.

RPRE1	0.3721925
RPRE2	0.3569996
RPRE3	0.4509193
RPRE4	0.5757129
RPRE5	0.121755
RPRE6	0.4866203
RPRE7	0.1858031
RPRE8	0.5438781
RPRE9	0.5086687
RPRE10	0.3585055
RPRE11	0.4368037
RPRE12	0.4139475
RPRE13	0.3807406
RPRE14	0.4719481
RPRE15	0.4638437

**Session 2: Thursday**

This will be a hands-on workshop where participants will work in pairs in a computer lab working with some of the data discussed in today's session. In particular, we will use two software packages (Excel and SPSS) to perform item analyses on some of the data used for placement in order to see how effective the items are in separating proficient from non-proficient students.

Session 3: Friday

In the first part of this session, we will work in small groups either designing an evaluation plan or developing assessment instruments. In the second part, each of the groups will present their designs for discussion by the whole group.