Assessment of Quantitative Reasoning to Enhance Educational Quality

Paper presentation at the annual

American Educational Research Association meeting

April, 2003

Chicago, Illinois

Donna L. Sundre,

Center for Assessment and Research Studies

James Madison University

Acknowledgements: This work could not have been conducted without the dedicated work of many, many faculty members representing several colleges of the University and three graduate students in the Assessment and Measurement masters and doctoral programs: Amy Thelk, Robin RiCharde, and BJ Miller.

The 2003 AERA theme: Accountability for Educational Quality: Shared Responsibility provides a solid foundation for research on the means by which quality assessment can improve both instructional delivery and student learning. A counterpoint to the prevalent accountability focus is intentionally sought to promote participation by both faculty and students toward assessment methods that can improve teaching and learning. This paper provides a progress report on an effort to enhance assessment of quantitative reasoning with an eye toward greater student engagement and assessment results that might better inform pedagogy.

Introduction

Calls for accountability continue to be heard, however the voices of educators and learners are difficult to discern over those of policy makers and testing companies. Significant educational change in classrooms and in students cannot take place through mandated testing or enhanced teacher knowledge (Brookhart, 2002). This paper is not about accountability; it is about seeking educational quality and sharing responsibility for educational quality with our students and our profession. As many authors have discussed, requiring mathematics courses in high school and college has not resulted in citizens with quantitative literacy (Hughes-Hallett, 2001). What seems to be necessary is contextualized instruction as part of a program of study (Davidson & McKinney, 2000; The Mathematics Association of America, 1998). Colleges and universities have been urged to take responsibility for monitoring programs of quantitative literacy through regular assessment activities that will guide program improvements (The Mathematics Association of America, 1998). However, mandates and legislation for assessment do not provide the assessment methods or infrastructure for conducting good assessment, and accountability is not assessment.

Many statewide governing boards of higher education, and certainly K-12 policy makers are mandating assessment of quantitative reasoning and similar constructs. While everyone

agrees on the importance of monitoring such an important construct, there is little agreement on its definition. Without consensus on what quantitative reasoning is, there can be no agreement on appropriate methods to measure it. This represents the first assessment challenge: construct definition. Second, the term accountability suggests provision of data for external stakeholders as a form of audit. The notion of a shared responsibility for assessment with the goal of enhancing teaching and learning would seem to mandate that the prime recipients of assessment processes and results be educators and learners. This represents a second assessment challenge: accountability vs. assessment. This paper provides a progress report on efforts to address both challenges: 1) to collectively define; and 2) assess quantitative reasoning at a midsize comprehensive institution of higher education for the purpose of providing information that informs pedagogy and program improvement.

Literature Review

There appears to be general agreement that there is a need for greater quantitative literacy (National Council on Education and the Disciplines, 2001; National Research Council, 1989) for our citizens. There have been efforts to provide some guidance on how the basic elements and skills involved might be expressed (National Council on Education and the Disciplines, 2001), and it is generally believed that enhanced educational quality might be achieved via capable instruction that was coupled with worthy assessment tailored to meaningfully inform teachers and students about their mutual progress and areas of weakness. However, discussions of the importance of the construct, whether it be labeled quantitative literacy, numeracy, quantitative reasoning, or statistical reasoning far surpass those pertaining to assessment. A paper by Thelk (2002) describes an attempt to locate through the Mental Measurement Yearbooks any instruments purporting to measure ‘quantitative reasoning’ or ‘mathematical reasoning’ at either the test or subscore level for postsecondary aged students. Her search resulted in only three

published instruments, and these provided only subscores. Apparently, no published instrument has been prepared to measure quantitative reasoning for postsecondary students. Thelk (2002) concluded that test users interested in assessing quantitative reasoning would most likely need to develop their own instrument. Fortunately, other researchers had discussed the importance of assessment in the service of instructional improvement and responded.

Assessment Focused on Instructional Improvement

Joan Garfield (1998) presented the instrument that she and Cliff Konold developed, the Statistical Reasoning Assessment (SRA). This instrument represents a welcome step forward in the design of instructional-friendly assessment tools. The scoring of responses was designed to provide two types of scores: correct reasoning and common misconceptions. The very notion of scoring misconceptions as well as correct reasoning represented a major step forward in informing instruction and providing meaningful feedback to students. Review of the items comprising the instrument revealed many meaningful concepts that faculty members at our institution strongly endorsed. While the instrument generated considerable interest, Garfield had cautioned that the reliability and validity results to date were not impressive. Liu (1998) also used the SRA and reported test-retest reliability of .70 for the correct total score and .75 for the incorrect reasoning scores. Garfield (1998) further suggested that only a small subset of reasoning strategies was being assessed. She encouraged continued progress toward development of sound instruments for research and evaluation studies.

Given the initial favorable review by selected faculty members at our institution, we took up the challenge. We also believed that the incorporation of correct reasoning and further identification of common misconceptions in reasoning would be powerful tools to aid assessment and instructional work. We have found faculty particularly eager to engage in these very labor-intensive efforts when they felt the work might result in more informative assessment results.

This paper provides progress in the development of a Quantitative Reasoning Quotient (QRQ) instrument designed to help define the construct further and to assess student progress in development of correct reasoning skills and competencies.

Revision of the Statistical Reasoning Assessment

The Quantitative Reasoning Quotient (QRQ) assessment method is a revision of Garfield's (1998) 20-item Statistical Reasoning Assessment, which was keyed to assess 8 correct reasoning skills and 8 misconceptions. We took Garfield's (1998) advice and attempted to modify the instrument to alleviate a few limitations: 1) low internal consistency; 2) item format and scoring omitted potentially important information; 3) difficulty in scoring; and 4) the instrument assessed a subset of reasoning strategies and misconceptions.

To further improve the internal consistency of the instrument, it was slightly revised to create additional items from existing alternatives. Review of the instrument format suggested that requesting students to respond to each alternative of the original items could create additional items. Rather than asking students to check from a list of possible rationales those they considered important, each rationale was presented as an item and students were requested to indicate whether they agreed or disagreed with the reasoning. We reasoned that a student may be able to recognize from a list an example of correct reasoning, but this would not eliminate the possibility that they might also endorse a common misconception when reviewing the same incident. By forcing students to respond to each alternative, we were able to create additional items. Adding to the sample of items from the construct domain is a sure way of enhancing the consistency or reliability of measurement. At the same time, we believed that this would retain potentially useful information that was not collected in the original item scoring that could contribute to both misconception and correct reasoning scores. Finally, we reasoned that when a

student disagreed with a misconception, this could be scored as evidence of correct statistical reasoning.

The scoring procedures for the original SRA could only be conducted by hand and were fairly labor intensive. We wanted to have an instrument that could be administered to large numbers of students and scored via computer. The modifications described earlier helped a great deal in making this possible. We were able to revise all items so that students could place their responses on machine-readable answer sheets. When scanned, the responses could be scored using SPSS or SAS programming languages. These modifications also rendered estimation of internal consistency reliability more feasible. Finally, faculty members reviewed all items, the keyed correct reasoning responses and all incorrect alternatives in an attempt to identify additional quantitative reasoning and misconception components. After several reviews by two faculty teams and a validation of these judgments, we had a 43-item instrument. Our faculty teams had identified 11 quantitative reasoning skills and 15 quantitative misconceptions and skill deficiencies. They had also keyed response alternatives to these categories. Table 1 provides a listing of the quantitative reasoning skills and the keyed items assessing them. Table 2 provides a parallel table for misconceptions and skill deficiencies. We affectionately called the instrument the Quantitative Reasoning Quotient (QRQ), because we hoped to someday form a quotient of correct reasoning over misconceptions over time.

Table 1. Quantitative Reasoning Skills Assessed by the QRQ.

Quantitative Reasoning Skills	Items Assessing:
C1 Correctly interprets probabilities	3d,27a
C2 Correctly interprets measures of central tendency	2d, 26b, 36b, 37b, 38a, 39b, 40b
C3 Understand how to select appropriate average	1d, 4a/b
C4 Correctly computes probability	11c, 41b, 42a, 43b
C5 Understand independence	12e, 13b, 14b, 15a, 16a, 17b, 18a,19e
C6 Understands sampling variability	9b, 28b, 30d, [30c=1 point]
C7 Distinguishes between correlation and causation	32b, 33a, 35b
C8 Correctly interprets two way tables	5a, 6d
C9 Understands the importance of large samples	8a, 22b, 29,
C10 Understands sources of bias and error	10a, 20a, 23b, 24a, 25a
C11 Recognizes features of good experimental Design	7b, 21b, 31b, 34b

Table 2. Quantitative Reasoning Misconceptions and Skill Deficiencies Assessed by the QRQ

Misconceptions and Skill Deficiencies	Items Assessing
M1 Misconceptions involving averages	1a, 26a, 30b/f, 36a, 37a, 38b, 40a
M2 Outcome orientation misconception	2e, 3a/b, 12a/b/d, 15b, 19a/b/d, 29c
M3 Good samples have to represent a high percentage of the population	21a, 22a, 31a, 34a
M4 Law of small numbers	28c, 29a
M5 Representativeness misconception	8b, 9a, 12c,13a, 14a,16b,17a,18b,19c
M6 Correlation implies causation	32a, 33b, 35a
M7 Equiprobability bias	27c, 41a, 42d, 43d
M8 Groups can only be compared if they are of the same size	7a
M9 Failure to distinguish the difference between a sample and a population	4c/d
M10 Failure to consider and evaluate all of the data	6a, 30a/e, 39a
M11 Inability to create and evaluate fractions or percents	6b/c, 11a/b
M12 Only large effects can be considered meaningful	[5b,6d]
M13 Failure to recognize potential sources of bias and error	10b, 20b, 23a, 24b, 25b
M14 Assumes more decimal places indicate greater Accuracy	1b
M15 Inability to interpret probabilities	2a/b/c, 3c/e, 27b, 28a, 30e, 41c/d, 42b/c 43a/c

Method

All data collection efforts took place during regularly scheduled University-wide Assessment Day activities. There are two formal Assessment Days scheduled on our campus each year. The first occurs in August just prior to the beginning of fall classes. Every entering first-year student is required to participate in a four-day orientation. One of the scheduled orientation activities is a three-hour assessment session. On the basis of the last two digits of their JMU ID, each student is assigned to a testing session and location. These assignments result in large, representative, and random groups of students assigned to a variety of assessment tasks. The second Assessment Day takes place in February of the spring semester. Classes for undergraduate students are canceled on this day, which results in no room conflicts and no time conflicts. Students with a cumulative credit hour total of 45-70 receive emails and are assigned to testing locations on the basis of the last two digits of the JMU ID. Since these numbers do not change, we are able to conduct repeated measures of the same students over time when our measures remain stable. We have outstanding participation with over 90% of students participating; failure to participate results in registration for classes being blocked until a make-up session has been completed. It should be noted that these particular assessment results bear no personal consequences for students and do not appear on their transcripts. This has resulted in quite a bit of study concerning examinee motivation in low-stakes conditions; our results suggest the vast majority of students put forth good effort on these Assessment Day tasks. Many academic departments use the spring Assessment Day to gather data from their graduating seniors. Every academic program annually collects and reports on their assessment data.

The QRQ was administered to large samples of students at two Assessment Days: the first during the spring 2002 Assessment Day and the second on the fall 2002 Assessment Day

activities. The QRQ was administered to 804 sophomore-level students during the spring 2002 Assessment Day activities. These results will be described first.

Results

The spring 2002 data collection effort provided a very large data set. The data were submitted to an SPSS scoring program, and total and quantitative reasoning scales were calculated. Each item was scored for two points each. Faculty deemed one alternative as partially correct, and it was scored as 1 point. To allow comparisons across correct reasoning and misconception scores, the means were scaled to a 0-2 point range by dividing by the number of items contributing to each score. Table 3 provides the descriptive statistics for the total test and quantitative reasoning scales. The internal consistency for the QRQ Total score was .62. Table 4 provides the parallel results for the spring 2002 QRQ Misconceptions and Skill Deficiencies. These score means have also been scaled to a range of 0-2 points for comparison purposes. It should be noted that misconception scores may be underestimates since a student couldn't only indicate one of several possible keyed misconceptions for a given item.

Table 3. Spring 2002 Results for Correct Reasoning QRQ Scores

(Scale = 0 to 2 points)

	Quantitative Reasoning Skill	Items Assessing and Scaled Score	
C1	Correctly interprets probabilities	Mean= .57	(29% correct)
C2	Correctly interprets measures of central tendency	Mean= 1.37	(69 % correct)
C3	Understands how to select an appropriate average	Mean= .80	(40% correct)
C4	Correctly computes probability a. understands probabilities as ratios b. uses combinational reasoning	Mean= .82	(41% correct)
C5	Understands independence	Mean= 1.39	(69% correct)
C6	Understand sampling variability	Mean= .68	(34% correct)
C7	Distinguishes between correlation and causation	Mean= 1.30	(65% correct)
C8	Correctly interprets two way tables	Mean= .51	(51% correct)
C9	Understands importance of large samples	Mean= 1.24	(62% correct)
C10	Understands sources of bias and error	Mean= 1.53	(77% correct)
C11	Recognizes features of good experimental design	Mean= 1.25	(62% correct)
	TOTAL N = 804 sophomores	*43 items (possible total of 86 pts)* Mean = 1.16	(60% correct) α =.62

Table 4. Spring 2002 Results for Misconceptions and Skill Deficiencies QRQ Scores

	Objective Assessed	Items Assessing
M1	Misconceptions involving averages	Mean= .66 (33%)
M2	Outcome orientation misconception	Mean= .39 (20%)
M3	Good samples have to represent a high percentage of the population	Mean= .74 (25%)
M4	Law of small numbers	Mean= .64 (32%)
M5	Representativeness misconception	Mean= .61 (31%)
M6	Correlation implies causation	Mean= .68 (34 %)
M7	Equiprobability bias	Mean= 1.19 (60%)
M8	Groups can only be compared if they are of the same size	Mean= 1.04 (52%)
M9	Failure to distinguish the difference between a sample and a population	Mean= .92 (46%)
M10	Failure to consider and evaluate all of the data	Mean= .69 (35%)
M11	Inability to create and evaluate fractions or percents	Mean= .38 (19%)
M12	Only large effects can be considered meaningful	Mean= .15 (8%)
M13	Failure to recognize potential sources of bias and error	Mean= .45 (23 %)
M14	Assumes more decimal places indicate greater accuracy	Mean= .05 (3%)
M15	Inability to interpret probabilities	Mean= .25 (11%)
	TOTAL	*43 items (possible total score of 86 points)*
		Mean= .76 (39 %) α= .62

Faculty Review of Spring 2002 Results

Faculty teams carefully reviewed these results. They expressed a desire that all alternatives be keyed back to either a correct reasoning category or a misconception. These faculty members had already identified a few new misconception categories and reviewed the items comprising the test to alleviate confusing items and modify alternatives to more closely fit with their misconceptions. They also determined that not all errors could be attributed to misconceptions in thinking, a few were more aptly described as skill deficiencies. This work resulted in a modified QRQ that was prepared for administration to entering first-year students in fall 2002.

A Revised Quantitative Reasoning Quotient Test

The new instrument was comprised of 40 multiple-choice items that assessed the same 11 correct quantitative reasoning scales and 15 misconceptions and skill deficiencies as those assessed with the spring 2002 instrument. Each item was scored at 2 points each. Again, one partially correct alternative was scored for 1 point. Table 5 provides the quantitative reasoning skills assessed by the revised QRQ, and Table 6 presents the misconceptions and skill deficiencies assessed. This new scoring mechanism provided 56 possible ways to achieve a misconception score. This is possible because alternatives from a single item might be coded to several misconceptions. However, the total possible score for both correct reasoning and misconceptions was 80.

Table 5. Quantitative Reasoning Skills Assessed by the Revised QRQ.

Quantitative Reasoning Skills	Items Assessing:
C1 Correctly interprets probabilities	3d,23a
C2 Correctly interprets measures of central tendency	2d, 32b, 33b, 34a, 35b, 36b, 37b
C3 Understand how to select appropriate average	1d, 4a/b
C4 Correctly computes probability	10c, 38b, 39a, 40b
C5 Understand independence	11e, 12b, 13a, 14b, 15a, 16e
C6 Understands sampling variability	8a, 24b, 26d, [26c=1 point]
C7 Distinguishes between correlation and causation	28b, 29a, 31b
C8 Correctly interprets two way tables	5a, 6d
C9 Understands the importance of large samples	19b, 25b
C10 Understands sources of bias and error	9a, 17a, 20b, 21a, 22a
C11 Recognizes features of good experimental Design	7b, 18b, 27b, 30b

Table 6. Quantitative Reasoning Misconceptions and Skill Deficiencies Assessed by the Revised QRQ

Misconceptions and Skill Deficiencies	Items Assessing
M1 Misconceptions involving averages	1a/c, 26b/e, 32a, 33a, 34b,36a, 37a
M2 Outcome orientation misconception	2e, 3a/b, 11a/b/d, 16a/b/d, 25c
M3 Good samples have to represent a high percentage of the population	18a, 19a, 27a, 30a
M4 Law of small numbers	24c, 25a
M5 Representativeness misconception	11c, 12a, 13b,14a, 15b,16c
M6 Correlation implies causation	28a, 29b, 31a
M7 Equiprobability bias	23c, 38a, 39d, 40d
M8 Groups can only be compared if they are of the same size	7a
M9 Failure to distinguish the difference between a sample and a population	4c/d
M10 Failure to consider and evaluate all of the data	[5a, 6a], [5b,6a], 26a, 35a
M11 Inability to create and evaluate fractions or percents	[5b, 6b/c], [5a,6b/c], 10a/b
M12 Only large effects can be considered meaningful	[5b,6d]
M13 Failure to recognize potential sources of bias and error	8b, 9b, 17b, 20a, 21b, 22b
M14 Assumes more decimal places indicate greater Accuracy	1b
M15 Inability to interpret probabilities	2a/b/c, 3c/e, 23b, 24a, 38c/d, 39b/c 40a/c

Fall 2002 QRQ Results

The revised QRQ was administered during the fall 2002 Assessment Day activities to a total of 1,083 entering first-year students. The correct reasoning results for this administration are presented in Table 7, and the results for the misconceptions and skill deficiencies are provided in Table 8. The internal consistency estimate for the instrument dropped slightly to .55 for this administration. We generally see a drop in internal consistency estimates for incoming students; most of our instruments are designed for sophomore-level students. While the variability in scores for entering students is generally higher, we typically see drops in our alphas. We believe entering student responses include more random error than those of sophomore students who have experienced more relevant course work.

Faculty Review of Fall 2002 QRQ Results

Our faculty reviewed these results. In general, they were pleased with the items and the development of additional correct reasoning and misconception categories. They believed that while the instrument had not achieved the level of internal consistency we would like to see, we had made a good start. Clearly, additional items would be necessary to create reliable scores upon which inferences might be warranted. They had hoped to see meaningful differences between the scores of entering students when compared with those of sophomore-level students. More specifically, they hoped to see higher correct reasoning scores and lower misconception scores for sophomore-level students. They requested a set of graphs to be prepared that would plot the scores of the entering first-year students with those of sophomore level students. We know from many years of analyzing sophomore-level test results, that while all students tested during the spring have accumulated between 45-70 credit hours, for many of these students, their work in mathematics and the sciences is quite meager. We often analyze their results in relation

Table 7. Fall 2002 Results for Correct Reasoning QRQ Scores

(Scale = 0 to 2 points)

	Quantitative Reasoning Skill	Items Assessing and Scaled Score	
C1	Correctly interprets probabilities	Mean= .62	(31% correct)
C2	Correctly interprets measures of central tendency	Mean= 1.23	(62 % correct)
C3	Understands how to select an appropriate average	Mean= 1.19	(60% correct)
C4	Correctly computes probability c. understands probabilities as ratios d. uses combinational reasoning	Mean= .89	(45% correct)
C5	Understands independence	Mean= 1.32	(66% correct)
C6	Understand sampling variability	Mean= .85	(42% correct)
C7	Distinguishes between correlation and causation	Mean= 1.18	(59% correct)
C8	Correctly interprets two way tables	Mean= 1.02	(51% correct)
C9	Understands importance of large samples	Mean= .93	(46% correct)
C10	Understands sources of bias and error	Mean= 1.46	(73% correct)
C11	Recognizes features of good experimental design	Mean= .92	(46% correct)
	TOTAL N = 1, 083 Entering First-Year students	*40 items (possible total of 80 pts)* Mean = 45.76	(57% correct) α =.55

Table 8. Fall 2002 Results for Misconceptions and Skill Deficiencies QRQ Scores

	Objective Assessed	Items Assessing
M1	Misconceptions involving averages	Mean= .87 (43%)
M2	Outcome orientation misconception	Mean= .50 (25%)
M3	Good samples have to represent a high percentage of the population	Mean= .94 (47%)
M4	Law of small numbers	Mean= .70 (35%)
M5	Representativeness misconception	Mean= .54 (27%)
M6	Correlation implies causation	Mean= .80 (40 %)
M7	Equiprobability bias	Mean= 1.06 (53%)
M8	Groups can only be compared if they are of the same size	Mean= .82 (41%)
M9	Failure to distinguish the difference between a sample and a population	Mean= .91 (46%)
M10	Failure to consider and evaluate all of the data	Mean= .19 (10%)
M11	Inability to create and evaluate fractions or percents	Mean= .40 (20%)
M12	Only large effects can be considered meaningful	Mean= .41 (20%)
M13	Failure to recognize potential sources of bias and error	Mean= .48 (24 %)
M14	Assumes more decimal places indicate greater accuracy	Mean= .04 (2%)
M15	Inability to interpret probabilities	Mean= .26 (13%)
	TOTAL	*40 items (possible total score of 80 points)*
		Mean=32.53 (40 %) α= .55

to the extent to which they have completed their general education requirements in the Natural World. We were able to identify 227 of the sophomores as those that had completed their 10-12 credit hour requirement in mathematics and sciences. We plotted the scaled means for all of the correct reasoning scores. These means are plotted in Figure 1. We also plotted the scores for the misconceptions; these appear in Figure 2. We were fairly disappointed with our findings. We did not see the hoped for improvements of students that had more experiences that we thought would be relevant and instructive. In fact, on several occasions, our entering students performed more capably than did those with more experience in college and in mathematics and science courses.

We advised caution in over interpreting these findings. It could be that our measure had not achieved sufficient reliability to begin to even look for meaningful differences, particularly at the scale mean level. This observation is obvious. However, an alternative explanation was also quite apparent that we were witnessing the same rather disappointing results that many other researchers had experienced (Steen, 2001). Misconceptions appear to be quite stubborn. Our students may be exhibiting difficulty transferring information from one context to another. When we interviewed students during and after administration of the QRQ, they appeared to be putting forth good effort, and many students expressed interest in the item types.

Our faculty are very committed to continuing to work hard with assessment activities. We are very fortunate to have their assistance in reviewing, developing, and conducting the very laborious 'backward translations' of items to correct reasoning and misconception categories. We feel confident that the assignments of these items have been validly placed within appropriate categories, but we surely do not have sufficient numbers of them to make valid inferences as yet. Our final summary and recommendations follow.

Figure 1. Plot of Spring 2002 and Fall 2002 Correct Reasoning Scale Means

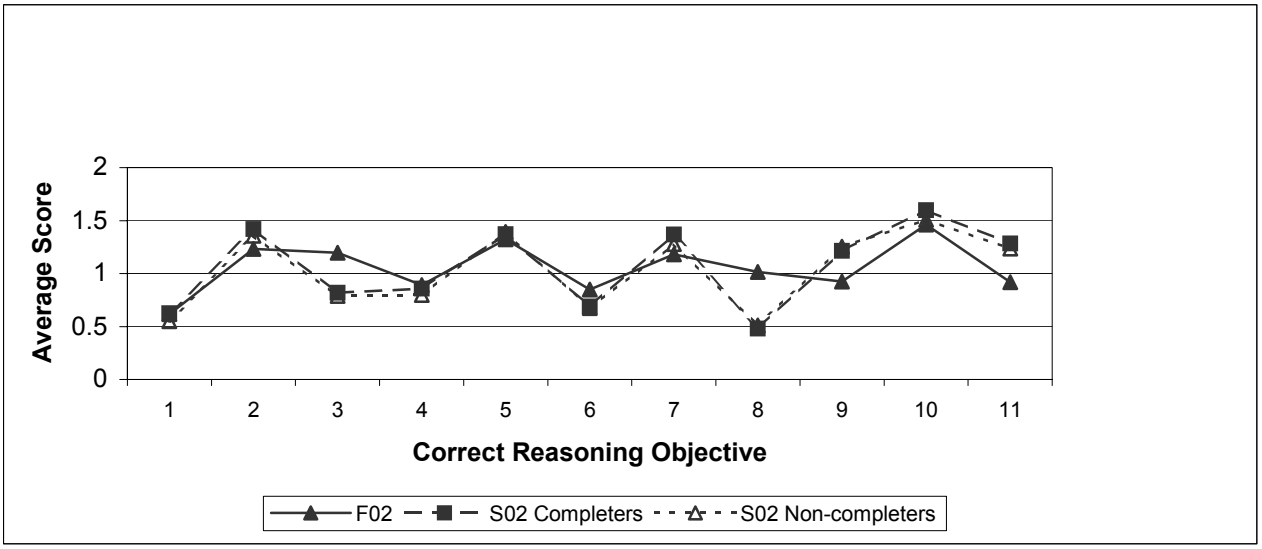
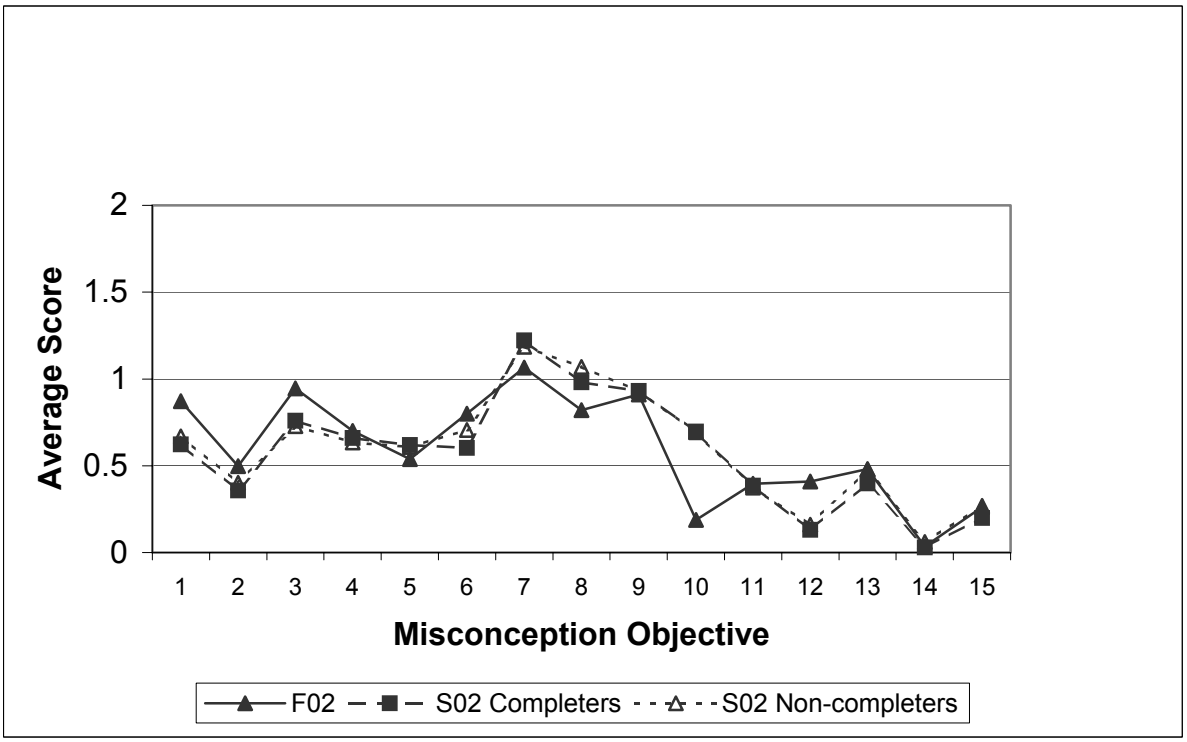


Figure 2. Plot of Spring 2002 and Fall 2002 Misconception Scale Means



What we learned

- Faculty members teaching in quantitative and scientific reasoning are very, very interested in receiving scores that indicate common flaws in thinking. They want to use information to enhance their instructional delivery, and they want to be able to see changes in student learning as a result. We have been very successful in this work in general scientific reasoning with other measures we have previously developed. However, our recent efforts with the quantitative reasoning domain were rather humbling..
- The process of backward translation is very time consuming and requires deep thought when the purpose is solely to classify items. Backward translation of item responses is much more arduous and many faculty were unable to complete the process. We need dedicated time to more fully involve faculty in these activities. Many faculty members realized that this work was important but simply could not carve out the time beyond their other duties to perform these tasks.
- Through the process of backward translation, we discovered that some of the misconceptions we identified were considered so important that many faculty believed we should have learning goals that address them. A good example of this awareness is the misconception that we called failure to recognize potential sources of bias and error. Faculty believed this misconception represented an important learning goal, and we incorporated it into our correct reasoning definition. Another example would be the development of the correct reasoning score for recognizes features of good experimental design. This work greatly enhanced our construct development and the validity of our instrument.

- Identification of misconceptions and correct reasoning are important for construct and instrument development, but making inferences about each of these will require many more items for reliable and valid interpretations of these scores.

Our work has shown that students show considerable interest in problems that are engaging and tasks that do not rely heavily on specific courses and topics they may not have studied. The assessment tasks we employed are general in nature, and therefore accessible to students; they feel like they have a chance to perform well regardless of the specific courses they have completed. They recognize many of the skills called upon as valuable.

Download of the Revised QRQ instrument and SPSS and SAS scoring programs will be available in April 2003 from:
http://www.jmu.edu/assessment

References

Brookart, S. M. (2002). What will teachers know about assessment, and how will that improve instruction? In R. W. Lissitz and W. C. Schafer (Eds.), *Assessment in educational reform.* (pp. 2-17). Boston , MA: Allyn & Bacon.

delMas, R. C., Garfield, J, and Chance, B.. (2001, January). Assessment as a means of instruction. Paper presented at the Joint Mathematics Meetings. San Diego, CA. Retrieved from http://www.gen.umn.edu/faculty_staff/delmas/jmm_2002/assess_instruct.html on 4/15/2002.

Ellis, W., Jr. (2001). Numerical common sense for all. In National Council on Education and the Disciplines (Ed.). *Mathematics and Democracy: The Case for Quantitative Literacy*. (pp. 61-66). United States: Author.

Garfield, J.B. (1998, April). *Challenges in Assessing Statistical Reasoning*. Paper presented at the meeting of the American Educational research Association, San Diego, CA.

Greeno, J. G., Pearson, P.D., and Schoenfeld, A. H. (1996). *Implications for NAEP of research on learning and cognition. Report of a study commissioned by the National Academy of Education.* Panel on the NAEP Trial State Assessment, Conducted by the Institute for Research on Learning. Stanford, CA: National Academy of Education.

Hughes-Hallett, D. (2001). Achieving numeracy: The challenge of implementation. In L.A. Steen (Ed.), *Mathematics and Democracy: The Case for Quantitative Literacy* (pp. 93-98). United States: Author.

Impara, J. C. & Plake, B. S. (Eds.). (1998). *The Thirteenth Mental Measurement Yearbook.* Lincoln, Nebraska: Buros Institute of Mental Measurements.

Liu, H. J. (1998). A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States. Doctoral dissertation, University of Minnesota, Minneapolis.

National Council on Education and the Disciplines. (2001). The Case for Quantitative Literacy. In L.A. Steen (Ed.), *Mathematics and Democracy: The Case for Quantitative Literacy* (pp. 1-22). United States: Author.

National Research Council. (1989). Everybody counts: A report to the nation on the future of mathematics education. Washington, DC: National Academy Press.

Steen, L. A. (Ed.) (2001). *Mathematics and democracy: The case for quantitative literacy.* National Council on Education and the Disciplines.

Sundre (2001). [Quantitative Reasoning Quotient Scale]. Unpublished assessment instrument. Center for Assessment and Research Studies. Harrisonburg, VA: James Madison University: Download at http://www.jmu.edu/assessment

Thelk, A. (2002). Seeking instruments to measure quantitative reasoning skills in postsecondary students: An overview. Center for Assessment and Research Studies. Harrisonburg, VA: James Madison University.