Numeracy
Advancing Education in Quantitative Literacy

Volume 3 | Issue 2 Article 2

2010

# Advancing Assessment of Quantitative and Scientific Reasoning

Donna L. Sundre
*James Madison University*, sundredl@jmu.edu
Amy D. Thelk
*James Madison University*, thelkad@jmu.edu



## Recommended Citation

Sundre, Donna L., and Amy D. Thelk. "Advancing Assessment of Quantitative and Scientific Reasoning." *Numeracy* 3, Iss. 2 (2010): Article 2. DOI: http://dx.doi.org/10.5038/1936-4660.3.2.2

# Advancing Assessment of Quantitative and Scientific Reasoning


## Abstract

Advancing Assessment of Quantitative and Scientific Reasoning is a four-year NSF Project (DUE-0618599) in part designed to evaluate the generalizability of quantitative (QR) and scientific reasoning (SR) assessment instruments created at James Madison University to four other four-year institutions with very distinct missions and student demographics. This article describes the methods, results, and findings we obtained in our studies. More specifically, we describe how to conduct content-alignment exercises in which faculty members map each item from a prospective test to the student learning objectives taught at the institution. Our results indicated that 92-100% of the QR and SR items were successfully mapped to each of the partner institutions' learning objectives. We also guided the partner institutions on assessing the balance of test items across the intended student learning objectives to assure greater content validity and coverage. The reliability (internal consistency) results from the partner institutions for the learning objectives and major subtests are strikingly similar across very different student populations. We interpret lower reliabilities from one institution to be the result of test administration and student motivation factors, the latter being a serious threat to the health and vigor of any assessment program. Validity study results at the partner institutions add to the evidence of construct validity of the QR and SR instruments. While our studies focus on QR and SR instruments, the methods will apply to other instruments and other institutions as they attempt to answer important questions about student learning outcomes.



## Keywords

quantitative reasoning, scientific reasoning, assessment, collegiate, instrument development, measurement, reliability, validity




## Cover Page Footnote


Donna Sundre has been with James Madison University's Center for Assessment and Research Studies (CARS) for over 20 years and, since 2003, its Executive Director. She is Professor of Graduate Psychology and frequently teaches Assessment Methods and Instrument Design. Dr. Sundre is an associate editor of this journal.

Amy Thelk is Director of Assessment and Evaluation at the College of Education, James Madison University. A graduate of JMU's Assessment and Measurement PhD program, Dr. Thelk's dissertation (2006) was *Examinee Awareness of Performance Expectation and its Effects on Motivation and Test Scores*.

# Introduction: Assessment of Quantitative and Scientific Reasoning in General Education

College students who do not major in math or science are generally exposed to these important disciplines through prescribed general university requirements and a small menu of general education courses. Although there is broad consensus that quantitative and scientific reasoning are critical for the future success of all students, there is little agreement on how to define these critical areas. There is even less agreement on how to assess these skills and competencies.

Despite an increasing demand for greater accountability, the general status of higher education assessment practice is not encouraging. Chun (2002) listed four methods used in higher education: actuarial, ratings of institutional quality, surveys, and direct measures of learning. He noted that it is disheartening to find that direct measures of student learning are the least systematically used of the four approaches. This is particularly discouraging because direct measure of student learning is the only methodology that should be used to guide improvements in curriculum and instruction.

Zemsky (2009) also commented on the lack of definitional clarity and availability of appropriate testing methods for assessing important student learning outcomes.

This article reports on results of a project designed to address these critical assessment needs. We hope to provide guidance on how to review instruments and better use results for program improvement.

## *Specific Assessment Issues*

Advancing Assessment of Quantitative and Scientific Reasoning is a four-year NSF Project (DUE-0618599) to further the development and dissemination of collegiate scientific and quantitative reasoning assessment tools. The project aimed to help address the nation's need for direct assessment of student learning in general education and more specifically to inform Science, Technology, Engineering and Mathematics (STEM) education. Without appropriate assessment methods, the nation will remain uninformed as to the growth and development of our students in quantitative and scientific reasoning. Such growth and development is a goal supported by every relevant learned society and espoused by every general education program across the nation.

In addition, the project attempted to directly address concerns delineated by the National Research Council in *Knowing what students know: The science and design of educational assessment* (Pelligrino et al. 2001). The NRC disputed the capacity of current assessments to measure complex knowledge and skills, provide information useful for teaching and improvement of learning, help us conceptualize how student understanding changes over time, and address the

important issues of fairness and equity. We have attempted to define quantitative and scientific reasoning and to develop items that assess these processes.

The project had six major objectives involving the home institution, James Madison University (JMU), and four partner institutions. The four partner institutions were Michigan State University (East Lansing, MI), St. Mary's University (San Antonio, TX), Truman State University (Kirksville, MO, and Virginia State University (Petersburg, VA). The major objectives were:

- Explore the psychometric quality and generalizability of the home institution's scientific reasoning (SR) and quantitative reasoning (QR) instruments to partner institutions having distinct missions and serving diverse populations.
- Develop scientifically based assessment plans to yield representative samples from the population and, through consultation and participation in a summer 2007 Faculty Institute, develop sound data collection plans at each of the partner institutions.
- Build assessment capacity at participating institutions through professional development in assessment practice, analytic methods, and data presentation to enhance curricular reflection and improvement.
- Create new assessment models and designs for adoption or adaptation by other institutions.
- Document potential barriers to effective assessment practice and explore solutions to the identified issues explored.
- Form scholarly communities of assessment practitioners in order to sustain the work at participating institutions and beyond.

In this paper, we focus on the first objective.

## History of the Test Instrument and Data Collection

The assessment instruments used in this project were developed by JMU's Center for Assessment and Research (CARS) and are available commercially through Madison Assessment LLC[1] of Washington DC. We used the ninth versions of the Quantitative Reasoning Test[2] (QR-9) and the Scientific Reasoning Test[3] (SR-9).

[1] http://www.madisonassessment.com/ (accessed June 12, 2010)
[2] http://www.madisonassessment.com/uploads/qr-9_manual_2008.pdf (accessed June 12, 2010). http://www.madisonassessment.com/assessment-testing/quantitative-reasoning-test/ (accessed June 12, 2010).
[3] http://www.madisonassessment.com/uploads/sr-9_manual_2008.pdf (accessed June 12, 2010). http://www.madisonassessment.com/assessment-testing/scientific-reasoning-test/ (accessed June 12, 2010).

By working collaboratively with STEM faculty, the CARS test developers have deliberately eliminated items we now refer to as "trivial pursuit," "factoids," or "basic skills mechanics" items. This type of item generally refers to recognition of specific course content and can readily be found in test item banks that accompany many published text books. Such items may be very appropriate for a quiz or examination for a given course but are not appropriate for assessment of general education objectives, which are much broader in scope.

An associated general rule that has informed the creation of our general education test items is that no item can privilege one course over another. Rather, we attempt to assess student ability to understand and use mathematics and science as ways of knowing. We believe this defines the heart of general education. We engaged our local STEM faculty in several summer item-writing workshops to guide them in following Cobb's (1998) principles in writing more innovative and interesting items that address higher levels of cognition.

We have conducted both quantitative and qualitative studies to gather information about item quality. For example, we interviewed students to determine which items they found confusing, intriguing, or interesting. We have conducted think-alouds with students to determine the strategies they used to solve problems (Thelk et al. 2006).

The QR and SR instruments developed at JMU have been successfully used for assessment of General Education program effectiveness in scientific and quantitative reasoning for over a decade. The exams have consistently shown improvement in their reliability estimates with each revision. Table 1 has a summary of results since 2001. This table clearly illustrates the consistent data collection efforts and the improvement of both instruments over time. The process employed in the development of the SR and QR follows that described by Wallace et al. (2009): we carefully identified and clarified the concept we were trying to measure, developed and fine-tuned the measurement over time, and engaged in formal testing of the instrument. To provide our faculty with quality assessment data, we ned quality instruments and credible samples of students.

The data supporting the results in Table 1 are generated from two Assessment Days conducted annually on the JMU campus. The first Assessment Day takes place in the fall semester just prior to the beginning of classes. All entering first-year students participate in this Assessment Day as an integral part of a required four-day orientation. Students are randomly assigned to classrooms on the basis of the last two digits of their student IDs, and each room has an assigned group of assessment tests. In other words, all students do not complete all assessment tests, but large random samples of students do complete each assessment. The second Assessment Day takes place on a Tuesday in mid-February. Classes are cancelled on this date, and all students with 45–70 credit hours (the midpoint of the undergraduate career) are again randomly assigned to rooms using the last two

digits of their student IDs. Because their ID numbers do not change, we can assure that students will retake the same instrument they were assigned upon entry. All students are required to participate, or their registration will be blocked. This Assessment Day is also used for data collection for graduating seniors for assessment in their majors. Our last fall Assessment Day involved over 4,000 entering students, and our spring Assessment Day includes over 3,500 participants. We have been using this data collection design for almost 25 years.

**Table 1**

**Number of Items, Sample Sizes and Reliability[1] for the Successive Forms of the Scientific and Quantitative Reasoning Tests (SR and QR), Fall 2000 through Fall 2009**

| Test Form[2] | | | | *First-year Students* | | | *Sophomores-Juniors* | | |
|---|---|---|---|---|---|---|---|---|---|
| | Items | | Semester | | SR | QR | | SR | QR |
| | SR | QR | | *N* | α | α | *N* | α | α |
| 5 | 27 | 23 | Fall 2000 | 994 | .54 | .50 | | | |
| | | | Spring 2001 | | | | 978 | .65 | .58 |
| | | | Fall 2001 | 746 | .56 | .52 | | | |
| | | | Spring 2002 | | | | 801 | .69 | .60 |
| | | | Fall 2002 | 1084 | .61 | .50 | | | |
| | | | Spring 2003 | | | | 1174 | .67 | .59 |
| 6 | 57 | 44 | Fall 2003 | 1304 | .75 | .64 | | | |
| | | | Spring 2004 | | | | 902 | .84 | .75 |
| 7 | 65 | 30 | Fall 2004 | 839 | .77 | .68 | | | |
| | | | Spring 2005 | | | | 770 | .83 | .75 |
| 8 | 50 | 24 | Fall 2005 | 1117 | .73 | .64 | | | |
| | | | Spring 2006 | | | | 510 | .82 | .73 |
| | | | Fall 2006 | 1186 | .76 | .63 | | | |
| | | | Spring 2007 | | | | 769 | .80 | .70 |
| 9 | 49 | 26 | Fall 2007 | 1408 | .71 | .64 | | | |
| | | | Spring 2008 | | | | 1020 | .74 | .66 |
| | | | Fall 2008 | 1592 | .80 | .66 | | | |
| | | | Spring 2009 | | | | 1113 | .83 | .70 |
| | | | Fall 2009 | 1408 | .78 | .64 | — | — | — |

1 Cronbach's alpha (α)

## *Preliminary Evidence of the Generalizability of the Instruments*

Although JMU has been approached by many institutions about using these instruments for general education assessment at their institutions, a primary concern was whether items developed to assess JMU learning objectives could be matched to the goals and objectives of other institutions. For existing instruments such as the SR and QR, the back-translation exercise (Dawis 1987) requires subject-area experts to review each item of the test to determine if it can be assigned to the learning objective it purports to assess. The individual content specialists then convene and compare their item-objective assignment decisions (Anderson et al. 2005).

Prior to the current project, JMU conducted two content-alignment workshops with two external clients (a community college system and a research university). Faculty content experts were asked to review each test *item by item* to determine alignment with their home learning objectives. Faculty from the first external site matched 76% of the JMU test items to *their own objectives*. Of equal importance, faculty members adopted one of JMU's General Education objectives after discovering that items they valued did not match any of their existing objectives. In other words, faculty from this external site discovered that the domain they were testing was underrepresented and elected to adopt one of JMU's learning objectives. At the second external site (the research University), faculty members matched 84% of JMU's QR and SR test items to their *home learning objectives*. Similar to faculty at the first site, they also discovered that JMU had included an objective that they had overlooked; they chose to adopt this new objective and all items mapping to it. These research results were reported by Sundre and Miller (2005) and strongly support the prudence of content-alignment exercises for test-selection activities. Both institutions continue to use the aforementioned tests.

A second set of content-alignment studies conducted with JMU faculty led to the identification of an improved methodology which we applied in the current project. This new technique, described by Miller et al. (2007), involved asking judges to review test items for alignment to student learning objectives one objective at a time (*objective by objective)*. The traditional method requires raters to assign items to objectives one item at a time (*item by item*); raters typically start with item one and attempt to locate an objective that the item seems to assess. They then move on to the next item and continue to the end of the test. Despite the fact that raters are encouraged to assign items to multiple learning objectives, they rarely do. Miller et al. (2007) demonstrated that asking faculty to consider only one objective at a time and to make dichotomous decisions (yes or no) as to whether each item measures an objective or not was: (1) less mentally taxing; (2)

actually took less time; and (3) produced a more dependable measurement design as assessed using Generalizability Theory (Shavelson and Webb, 1991).

Overall, the results of these two sets of studies were very satisfying and speak to the congruence of our items to the scientific and quantitative reasoning objectives of educational institutions with very different missions (a community college system vs. a research institution). They also provided a strong framework for use of the new content-alignment procedure with new partners. We built upon these successful experiences with our four external partners.

## *The Value of Content Alignment*

The first part of an instrument review should include careful consideration of content alignment of test items to stated student learning objectives (Miller et al. 2007). We have found that engaging the faculty who teach in the content area in the instrument selection process is very worthwhile. Faculty involvement in test selection and content alignment has produced several highly desirous outcomes: (1) they have much better understanding of the institution's stated learning objectives; (2) they can attest to the fit of the selected instrument to those objectives; (3) they have much greater confidence and interest in the assessment results; and (4) their capacity to actually use the assessment results to improve their curricular coverage and instructional intensity also improve. Faculty members are now much more willing to make an inference concerning whether or not student learning has occurred. This highlights the difference between a survey of opinions and true student learning assessment.

The content-alignment technique is an example of using assessment as a strategy to improve learning. More specifically, the emphasis is on improvement of learning over simply reporting data, and using information gathered via assessment to inform programming and decision-making at the institutional level.

When an institution is able to map a high percentage of test items to its goals and objectives, early evidence for generalizability of the instrument exists. Observing high percentages of items successfully aligned provides support for the content validity of the instrument. Our partner institutions, using the *objective-by-objective* content-alignment method at our Summer Institute were able to map between 61 and 66 of the JMU items (92% to 100% of the total number of items) to their *home institution learning objectives*. These were our most positive results to date. All of the partner institutions left with a deeper appreciation of the instruments' suitability for their general education programs.

Keep in mind, however, that mapping of items alone is not sufficient—balance across objectives must be obtained as well. If a team found that there were few or no test items applicable to one of their objectives, the project design allowed for creation of additional items to assure balance across the learning objectives. This is a recommended test-review procedure for all programs

considering use of a new instrument: assure that the balance of items to your home institutions provides sufficient content coverage and balance. If there are not enough items to cover your objectives, writing additional items is an important activity.

## Test Data Results

As mentioned above, four of the five institutions have completed fall data collection. At this stage, reliabilities provide the most compelling generalizability evidence; a later phase of the project involved validity studies conducted at each of the partner institutions. Table 2 shows the reliabilities for each institution as mapped to the JMU objectives, QR and SR scores, and total score. Since the number of items mapping to the individual objectives is relatively low, the associated reliabilities are low. Until the reliabilities for the individual learning objectives are higher, we can only use the QR and SR scores to form inferences. We report the objective-level reliability estimates here for completeness and to advise readers to seek similar information prior to using objective-level data as a research variable. Note that the means are *not* provided. This project was not intended to promote comparison of students across institutions.

Review of Table 2 reveals fairly consistent reliability results, particularly for the QR and SR scores. In general, the observed reliabilities for VSU appear a bit lower than the other institutions, and we believe this is due to administrative constraints. As noted in the table, this institution was compelled to gather data using a course-embedded technique that spanned two class occasions. This procedure led to an inordinate amount of missing data; the team leaders also suggested that many students did not appear motivated to complete the tasks. This should serve as a caution to institutions; while none of the institutions in the study were using the QR and SR in a testing context for which personal consequences would be in evidence (high-stakes testing), only this institution reported examinee motivation issues that they felt seriously impacted student performances. Low-stakes assessment conditions are known to influence both student motivation and performances; therefore, attention to administrative detail is paramount. At JMU, we have dedicated considerable time and effort to the study of examinee motivation in low-stakes testing conditions. Our Motivation Research Institute[4] which operates within JMU's Center for Assessment and Research Studies is devoted to research associated with student and examinee motivation. Publications and presentations are listed at this site, and most are downloadable. Interested readers may also wish to review a special issue of the *Journal of General Education* (2009, Vol. 58, Number 3) that focuses on examinee

[4] http://www.jmu.edu/assessment/research/MRI_Overview.htm (accessed June 3, 2010)

motivation research and solutions. All other institutions had assessment procedures in place that communicated institutional commitment to the data collection and the importance of the findings.

**Table 2**

**Sample Sizes, Context, and Reliabilities[1] for the Four NSF-Project Partner Institutions as Mapped to JMU Objectives**

**Sample and Context**

JMU – First-year students, tested immediately prior to the first semester, in one testing session.
SMU – Full-time, first-time freshmen were tested in one session, on a walk-in basis.
TSU – Juniors were tested as part of regular annual testing activity for that group
VSU – First-year students were tested in Freshman Studies course sections. Test was given over two 45-minute sessions.

| **Objectives** | **JMU** ***N*=1408** | **SMU** ***N*=426** | **TSU** ***N*=345** | **VSU** ***N*=653** |
|---|---|---|---|---|
| ❑ JMU1: Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudo-science. | α = .43 | α = .41 | α = .39 | α = .23 |
| ❑ JMU2: Use theories and models as unifying principles that help us understand natural phenomena and make predictions. | α = .20 | α = .28 | α = .33 | α = .21 |
| ❑ JMU3: Recognize the interdependence of applied research, basic research, and technology, and how they affect society. | α = .47 | α = .45 | α = .64 | α = .40 |
| ❑ JMU4: Illustrate the interdependence between developments in science and social and ethical issues. | α = .25 | α = .34 | α = .19 | α = .12 |
| ❑ JMU5: Use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomenon. | α = .58 | α = .55 | α = .63 | α = .48 |
| ❑ JMU6: Discriminate between association and causation, and identify the types of evidence used to establish causation. | α = .45 | α = .43 | α = .27 | α = .31 |
| ❑ JMU7: Formulate hypotheses, identify relevant variables, and design experiments to test hypotheses. | α = .59 | α = .60 | α = .47 | α = .57 |
| ❑ JMU8: Evaluate the credibility, use, and misuse of scientific and mathematical information in scientific developments and public-policy issues. | α = .32 | α = .25 | α = .24 | α = -.07 |
| Quantitative Reasoning (QR) Objectives 5 & 6 | α = .64 | α = .63 | α = .66 | α = .55 |
| Scientific Reasoning (SR) | α = .71 | α = .73 | α = .71 | α = .60 |
| Total | α = .78 | α = .79 | α = .77 | α = .71 |

1 Cronbach's alpha (α)

A few of our partner institutions have correlated QR and SR scores with those obtained from other nationally marketed instruments from ETS and ACT. The correlations (ranging from positive 0.35 to 0.55) provide support for concurrent validity. Truman State reported that QR and SR discriminate well between under- and upper-class students as well as science and mathematics majors vs. other majors. St. Mary's identified expected differences in entering students from different feeder high schools.

Over the years, we have conducted many studies at JMU exploring QR and SR test score validation. In the bulleted list below, we provide a summary of some of the research questions we have posed and answered via assessment analysis. These results provide compelling evidence, not only of the utility of this instrument, but also the efficacy of our general education program. Full assessment reports are available for download from JMU's General Education Web site.[5]

- Reliability estimates for both instruments are stable even with reduction in items; reliability is higher for sophomores than freshmen.
- Sophomores and juniors with 45–70 credits do not score differently from one another across academic years; however, sophomore samples consistently score significantly higher than entering freshmen.
- Scores on both instruments increase significantly with increasing numbers of related general-education courses completed.
- Multiple regression analyses reveal that related advanced-placement (AP) and JMU general-education courses both significantly predict SR and QR scores. In contrast, related transfer credits do not. Of additional interest, cumulative credit hours across subject areas negatively predict SR and QR scores. In other words, test scores are not enhanced via academic maturation through undifferentiated course taking; the tests are sensitive only to highly related course work.
- Over 90% of correlations between relevant course grades and scores on both instruments were positive (These correlations generally are in the 0.30–0.50 range).
- The Biology department uses the QR and SR tests as a supplemental assessment tool for their graduating seniors. Their students perform exceptionally better than sophomores and juniors who have completed their general education requirements.
- In recent years we have developed community standards established by faculty for student QR and SR test performances. This process has yielded some intriguing findings; we observe that about 75% of students meet or exceed faculty expectations upon completion of related course work.

[5] http://www.jmu.edu/assessment/JMUAssess/GenEdOverview.htm (accessed June 2, 2010)

Some objectives appear more difficult to master than others (Objective 6 [Table 2]: discriminating correlation and causation, for example). We also believe our faculty members have very high expectations.

Prior to this project, we had increasing evidence that important inferences we wish to make about student learning and development at our institution are valid, but the key question remained about whether such results could be generalized to other institutions. Findings to date lend support regarding the generalizability of the exam to other settings. Although the findings reported here are specific to the QR and SR instruments, readers may apply the framework for evaluating generalizability of any instrument.

# Discussion

This project addresses the assessment of an instrument's generalizability across institutions. There is little precedence for this type of work with postsecondary students in the quantitative and scientific reasoning domain. In fact, Chun (2002), Klein (2002), and Zemsky (2009) have all bemoaned the dearth of meaningful definitions, tests, and reported results across higher education. This project has provided meaningful information concerning the generalizability of the test items to the QR and SR learning objectives of four partner institutions. Further, the project has also demonstrated the stability of the reliability estimates for the QR and SR scores across four very distinct institutions of higher education. This project is now poised to move forward with validity evidence from the partner institutions. Each institution developed research questions they intended to pose and answer in the next phase of the project. Stay tuned for results.

By administering this test as consistently as possible across institutions, the value of regular assessment can begin to be showcased. Evaluation of programs and student learning can, and should, occur on a regular cycle. By incorporating regular assessment into the annual rhythm on campus, the process goes from being burdensome and inconvenient to expected and efficient. Since JMU has been in the practice of student-learning assessment for two decades (and this exam in particular for over ten years) the historic information we bring to the project eases the partner institutions' responsibilities of explaining and interpreting the instrument and convincing the stakeholders of the worth of regular assessment.

JMU has invested over ten years in a significant, long-term interdisciplinary collaboration by which scientific and quantitative reasoning objectives have been carefully crafted, reviewed, and revised. Through collaborative work, our interdisciplinary team has provided credible evidence to support the scientific and quantitative reasoning objectives we have crafted, the instruments we have developed, the assessment practices we model, and the reporting strategies we

have employed. JMU just received notification that the QR and SR component of our General Education program has been selected as the sole recipient of the 2009 Association for General and Liberal Studies (AGLS) award for Improving General Education, in part because of the efforts to use assessment data for making improvements in the courses offered.

## Concluding Remarks

We have growing evidence that our assessment instruments and our enthusiasm for assessment will generalize to other institutions in need of sound assessment methods and practices. Such instruments and practices are sorely needed by institutions, researchers, collegiate instructors, and other funded projects. This project provided the opportunity to assess the instruments' generalizability to institutions serving a wider variety of missions, to help explore and present new models of assessment practice that other institutions can adopt or adapt for their own use, and to directly assess the viability and validity of the instrument's use with underrepresented students.

We believe that we can promote professional development and build institutional capacity to engage in quality assessment practice. This project has and will continue to enhance the sustainability of assessment work and collaboration on each campus far beyond grant funding. The development of scholarly and truly interdisciplinary communities within and across institutions will directly contribute to new research on teaching and learning that can impact the field. Through the formation of partnerships with the participating institutions, and thanks to NSF funding, we believe these lofty objectives so central to the assessment of student scientific and quantitative achievement will be achieved.

## Acknowledgments

The authors would like to acknowledge the National Science Foundation (Award # 0618599) for the funding that propelled this project from a dream to a reality. In addition, the contributions of James Madison University, the CARS, the JMU STEM faculty members who labored over learning objectives and wrote examination items, and countless graduate students who provided committed professional-level work toward the development and revision of the SR and QR instruments over the past decade must be recognized. Finally, gratitude is expressed to the team members from all of the partner institutions for their steadfast support and scholarly energy. We thank them all.

## References


Anderson, R. D. and A. D. Thelk. 2005. The back translation: A good practice in instrument selection. *Assessment Update*, 17(2): 14-15.

Chun, M. 2002. Looking where the light is better: A review of the literature on assessing higher education quality. *Peer Review*, 4(2/3): 16-25.

Cobb, G. W. 1998. The objective-format question in statistics: Dead horse, old bath water, or overlooked baby? Invited paper presented to American Educational Research Association. San Diego, CA.

Dawis, R. 1987. Scale construction. *Journal of Counseling Psychology*, 34: 481–489.

Klein, S. 2002. Direct assessment of cumulative student learning. *Peer Review*, 4(2/3): 26-28.

Miller, B. J., C. Setzer, D. L. Sundre and X. Zeng. 2007. Content validity: A comparison of two methods. Paper presentation to the National Council on Measurement in Education. Chicago, IL.

Pellegrino, J., N. Chudowsky, and R. Glaser, eds. 2001. *Knowing what students know: The science and design of educational assessment*. National Research Council. Committee on the Foundations of Assessment. Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences Education. Washington, DC: National Academy Press.

Shavelson, R.J., and N.M.Webb. 1991. *Generalizability theory: A primer.* Thousand Oaks, CA: Sage Publications, Inc.

Sundre, D. L. and B. J. Miller. 2005. Continued refinement of an assessment instrument: JMU's scientific and quantitative reasoning tests. Paper presented at the annual meeting of the Virginia Assessment Group. Virginia Beach, VA.

Thelk, A. D., and E. R. Hoole. 2006. What Are You Thinking? Postsecondary Student Think-Alouds of Scientific and Quantitative Reasoning Items. *Journal of General Education*, (55)1: 17-39.

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 (accessed June 12, 2010).

Zemsky, R. 2009. The to-do list. *Inside Higher Ed* (September14). http://www.insidehighered.com/views/2009/09/14/zemsky (accessed June 12, 2010).