

Advancing Assessment of Quantitative and Scientific Reasoning

Donna L. Sundre

Amy D. Thelk

James Madison University

Abstract

This NSF-funded research project spans three years and utilizes five higher-education institutions. While all involved are four-year institutions, the demographic characteristics of the students attending each school vary. The purpose of the research is to evaluate the generalizability of a quantitative/scientific instrument created at James Madison University. The process of content alignment is presented, along with its relevance to this research. Test reliabilities from each site are presented and discussed.

Introduction

Although general education is integral to the attainment to scientific and quantitative literacy in postsecondary students, the lack of available instrumentation and methodologies remains. There is consensus that these skills are critical, but there is little agreement regarding definitions or assessment methods (Chun, 2001). Chun listed four assessment methods used in higher education: actuarial, ratings of institutional quality, surveys, and direct measures of learning. Direct measures are the least systematically used of the four approaches, yet these are the methods that provide the richest information about student growth and development and help us to modify instructional delivery to foster and maintain learning. Chun concluded that, “The key is to focus on developing better methods to directly assess student learning” (p.25). Hersh & Benjamin (2002) listed four barriers to assessing general education learning outcomes: confusion, definitional drift, lack of adequate measures, and the misconception that general education cannot be measured. This project addresses all of these concerns with special emphasis on the dearth of adequate measures.

Assessment of General Education. Despite the importance of general education to the attainment of scientific and quantitative literacy, the available methodologies remain insufficient. While there is agreement that these skills are important, there is little consensus what exactly this literacy encompasses, as well as how one would assess quantitative and scientific literacy (Chun, 2002). Chun listed four methods used in higher education: actuarial, ratings of institutional quality, surveys, and direct measures of learning. It is disheartening to find that direct measures are the least systematically used of the four approaches. Direct measures can best inform us about student growth and

development and help us to modify instructional delivery to foster and maintain learning. Chun concluded that, “The key is to focus on developing better methods to directly assess student learning” (p.25). Klein (2002) concurred by insisting that “development of appropriate measures of the quality of undergraduate education is the missing but essential ingredient needed to improve our understanding of the tradeoffs between access, productivity, and quality in American higher education” (p. 26). In addition, the nation’s students perceive their general education requirements as just that, a requirement to ‘get out of the way’ before they can take the courses that they think deem most important for employment or advanced degrees. However, recent research strongly suggests that the skills advanced by general education are among those most in demand by employers (National Research Council, 2001).

Specific Assessment Issues. Building upon 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, this project furthers the development and dissemination of collegiate scientific and quantitative reasoning assessment tools. 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 the general education program of every institution across the nation.

In addition, this project attempts to directly address the concerns delineated by the National Research Council (NRC) in their 2001 report, *Knowing what students know: The science and design of educational assessment*. 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. Our research is increasing the ability to assess STEM learning in general education across a diverse set of institutions and programs. Primarily, this has been achieved through a focus on explication of what general education is. Since students do not take the same courses, general education assessment cannot be a set of content-specific items. We have attempted to define quantitative and scientific reasoning and to develop items that assess these processes.

Project Objectives

This NSF-funded project has six major objectives involving the home institution and four partner institutions. First, the psychometric quality and generalizability of the scientific reasoning (SR) and quantitative reasoning (QR) instruments to partner institutions having diverse missions and serving diverse populations will be explored. Second, scientifically based assessment plans were developed. Through consultation and participation in a summer 2007 Faculty Institute, data collection plans were developed for adoption at the NSF partner institutions. Third, building assessment capacity at participating institutions through professional development in assessment practice, analytic methods, and data presentation to enhance curricular reflection and improvement was undertaken. Fourth, new assessment models and designs for adoption or adaptation by other institutions are being created. Fifth, potential barriers to effective assessment practice are being documented, and solutions to these issues explored. Finally, scholarly communities of assessment practitioners are being formed, in order to sustain work at

participating institutions and beyond. Specifically, this report focuses on exploring the psychometric quality and generalizability of James Madison University's Quantitative and Scientific Reasoning instruments to institutions with diverse missions and serving diverse populations.

History of the test instrument

For this research, the ninth versions of our two locally developed instruments that measure QR and SR are being employed. The Principal Investigators believe these instruments have important utility for many other institutions. Working collaboratively with our STEM faculty, we have learned a great deal about what general education is and how to create appropriate items. We have deliberately eliminated items we now refer to as 'trivial pursuit,' 'factoids,' or 'basic skills mechanics' items. A rule we try to follow 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 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 used to solve problems (Thek & Hoole, 2006). We engaged our local STEM faculty in several summer Faculty Institutes in which we guided them following Cobb's (1998) principles in writing more innovative and interesting items that address higher levels of cognition than the previous versions.

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.

Table 1
Sample Sizes and Reliability (α = Cronbach's alpha) for the Scientific and Quantitative Reasoning Tests (SR and QR), Fall 2000 through Spring 2008

Acad. Yr.	Test Form	Semester	<i>First-year Students</i>			<i>Sophomores-Juniors</i>		
			<i>N</i>	<i>SR</i> α	<i>QR</i> α	<i>N</i>	<i>SR</i> α	<i>QR</i> α
2000-2001	5	Fall 2000 Spring 2001	994	.54	.50	978	.65	.58
2001-2002	5	Fall 2001 Spring 2002	746	.56	.52	801	.69	.60
2002-2003	5	Fall 2002 Spring 2003	1084	.61	.50	1174	.67	.59
2003-2004	6	Fall 2003 Spring 2004	1304	.75	.64	902	.84	.75
2004-2005	7	Fall 2004 Spring 2005	839	.77	.68	770	.83	.75
2005-2006	8	Fall 2005 Spring 2006	1117	.73	.64	510	.82	.73
2006-2007	8	Fall 2006 Spring 2007	1186	.76	.63	769	.80	.70
2007-2008	9	Fall 2007 Spring 2008	1408	.71	.64	1020	.74	.66

Five separate versions of the SR and QR were administered during the academic years 2000-01 to 2007-08. The number of items comprising each of these tests is as follows:

Form 5: SR – 27; QR – 23

Form 6: SR – 57; QR – 44

Form 7: SR – 65; QR – 30

Form 8: SR – 50; QR – 24

Form 9: SR – 49; QR – 26

Preliminary Evidence of the Generalizability of the Instruments. The home site of this research has been approached by institutions that had learned about our assessment practices and results at conferences to purchase the instruments for use at their institutions. A primary concern was whether items we developed to assess our own objectives could be matched to their goals and objectives. 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 that it assesses. The individual content specialists then convene and compare their item-objective assignment decisions (Anderson & Thelk, 2005). We conducted two content alignment workshops with two external clients and included an important twist. Faculty from the first external site matched 76% of test items to *their own objectives*. Of equal importance, faculty members adopted one of the home site's General Education objectives after discovering that items they valued did not match any of their own objectives. In other words, faculty from this external site discovered that their construct was underrepresented and elected to adopt an additional learning objective. At the second external site, faculty members matched 84% of our items to their learning objectives. Similar to faculty at the first site, they also discovered that the home site had included an objective that they had overlooked; they chose to adopt this new objective and the items mapping to it. These results were reported by Sundre and Miller (2005). Both institutions continue to use the aforementioned tests. Our continued research identified an improved content alignment methodology (Miller, Setzer, Sundre & Zeng, 2007). These results were very satisfying and speak to the congruence of our items to the scientific and

quantitative reasoning objectives from two very different educational institutions. We are building upon these successful experiences with our four external partners.

We have conducted many studies exploring the validity of the test scores we have produced. These research procedures are consistent with the professional standards of the educational research, psychology, and measurement fields (AERA, APA, NCME, 1999). All items included in the test have been successfully mapped to the appropriate quantitative and scientific reasoning objectives by our STEM faculty and independent raters. These content alignment processes have provided support for the content validity of both instruments. Our local validity studies have included several additional methodologies.

We have identified students with different course-taking histories to determine if the number of courses taken in related general education courses impacts test performances. We have correlated student test performances with grades in science and mathematics courses. In addition, because we had the use of the fifth version of the instruments over several years, we were able to conduct a number of true repeated measures analyses (i.e., assessment of the same students with the same instrument as entering students and again as sophomores). It is important to note that these interpretive reports were generated through the collaboration of assessment and measurement experts working closely with STEM faculty. 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.

- 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 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.
- Over 90% of correlations between relevant course grades and scores on both instruments were positive.

Prior to the NSF 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 can be generalized to other institutions.

Results

At this writing, data had been collected at four of the five collaborating schools. Findings to date lend support regarding the generalizability of the exam to other settings. Although the findings reported here are specific to our research, readers may apply the framework for evaluating generalizability of any instrument.

Content alignment. The first part of an instrument review should include careful consideration of content alignment (Miller, Setzer, Sundre & Zeng, 2007). By initially aligning test items to the objectives of each institution's respective general education program, later findings inform the stakeholders whether or not learning has occurred. This highlights the difference between a survey of opinions and true student learning assessment. Performing the content alignment exercise also alleviates the tendency to collect and present data in a rote fashion, without attending to the meaning inherent in the findings. By gathering information on exactly how students perform on each segment of the domain, assessment results can inform subsequent programmatic decisions. For example, if student scores suggest that learning is taking place, then the program is meeting its goals. If the opposite appears to be true, then decision-makers can take this into account when planning number of course sections, considering changes in general education requirements, or reviewing syllabuses for related courses. The content alignment technique is an example of using assessment as a strategy to improve learning (specifically, the emphasis of 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. Table 2 exhibits the level of alignment.

Table 2

Percentage of items mapped for each NSF partner institution's own objectives

Institution	Percentage of items mapping to institution-specific objectives	Number of unmapped items
Truman State	100%	0
Virginia State	97%	2
St. Mary's	92%	5
Michigan State	98%	1

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, they created additional items.

Test data results. As mentioned above, four of the five partner institutions have completed fall data collection. At this stage, reliabilities provide the most compelling generalizability evidence; the next phase of the project involves validity studies. Table 3 shows the reliabilities for each institution as mapped to the JMU objectives, QR and SR subscores, and Total score. Since the number of items mapping to each objective is relatively low, the reliabilities at the objective level are not high enough to be used in practice. At JMU, we routinely report subscale-score and total-score reliabilities.

Table 3

Sample sizes, context, and reliabilities (Cronbach's alpha) for four NSF partner institutions as mapped to JMU objectives

	JMU N=1408	SMU N=426	TSU N=345	VSU N=653
Sample and Context:	First year students, tested immediately prior to the first semester, in one testing session.	Full-time, first time freshmen were tested in one session, on a walk-in basis.	Juniors were tested as part of regular annual testing activity for that group.	First year students were tested in Freshman Studies course sections. Test was given over two 45-minute sessions.
Objectives				
□ JMU1: Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudo-science.	$\alpha = .43$	$\alpha = .41$	$\alpha = .39$	$\alpha = .23$
□ JMU2: Use theories and models as unifying principles that help us understand natural phenomena and make predictions.	$\alpha = .20$	$\alpha = .28$	$\alpha = .33$	$\alpha = .21$
□ JMU3: Recognize the interdependence of applied research, basic research, and technology, and how they affect society.	$\alpha = .47$	$\alpha = .45$	$\alpha = .64$	$\alpha = .40$
□ JMU4: Illustrate the interdependence between developments in science and social and ethical issues.	$\alpha = .25$	$\alpha = .34$	$\alpha = .19$	$\alpha = .12$
□ JMU5: Use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomenon.	$\alpha = .58$	$\alpha = .55$	$\alpha = .63$	$\alpha = .48$

□ JMU6: Discriminate between association and causation, and identify the types of evidence used to establish causation.	$\alpha = .45$	$\alpha = .43$	$\alpha = .27$	$\alpha = .31$
□ JMU7: Formulate hypotheses, identify relevant variables, and design experiments to test hypotheses.	$\alpha = .59$	$\alpha = .60$	$\alpha = .47$	$\alpha = .57$
□ JMU8: Evaluate the credibility, use, and misuse of scientific and mathematical information in scientific developments and public-policy issues.	$\alpha = .32$	$\alpha = .25$	$\alpha = .24$	$\alpha = -.07$
Quantitative Reasoning (QR) Objectives 5 & 6	$\alpha = .64$	$\alpha = .63$	$\alpha = .66$	$\alpha = .55$
Scientific Reasoning (SR)	$\alpha = .71$	$\alpha = .73$	$\alpha = .71$	$\alpha = .60$
Total	$\alpha = .78$	$\alpha = .79$	$\alpha = .77$	$\alpha = .71$

Note that the means are *not* provided. This activity is not intended to promote comparison of students across institutions.

Discussion

This NSF 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. Considering the generalizability of the test assures a higher quality of assessment work, since the scores obtained at a given institution can be interpreted and reported with greater confidence. Addressing instrument suitability and performance in a variety of settings also begins to pave the way for allowing performance gaps to be addressed within or among schools.

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 10 years) the historic information we bring to the project eases the partner institutions' responsibilities of explaining/interpreting the instrument, and convincing the stakeholders of the worth of regular assessment.

The home site of this research has invested over 10 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. 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. These instruments and practices are sorely needed by institutions, researchers, collegiate instructors, and other funded projects. This NSF-funded 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 will 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.

References

- American Educational Research Association, American Psychological Association and the National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Anderson, R. D. and Thelk, A. D. (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: April.
- Dawis, R. (1987). Scale construction. *Journal of Counseling Psychology*, 34, 481-489.

- Hersh, R. H., & Benjamin, R. (2002). Assessing selected liberal education outcomes: A new approach. *Peer Review*, 4(2/3), 11-15.
- Klein, S. (2002). Direct assessment of cumulative student learning. *Peer Review*, 4(2/3), 26-28.
- Miller, B. J., Setzer, C., Sundre, D. L., & Zeng, X. (2007, April). *Content validity: A comparison of two methods*. Paper presentation to the National Council on Measurement in Education. Chicago, IL.
- National Research Council. (2001). *Knowing what students know: the science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser. (Eds.). Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences Education. Washington, DC: National Academy Press.
- Sundre, D. L. & Miller, B. J. (2005, November). *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., & Hoole, E. R. (2006). What Are You Thinking? Postsecondary Student Think-Alouds of Scientific and Quantitative Reasoning Items. *Journal of General Education*, (55)1.