4

This chapter describes recent attempts to measure context-rich application of quantitative reasoning. Initial evidence suggests that teaching to explicitly stated learning goals—whether in the context of a limited cluster of courses or in offerings across the curriculum—can shape student aptitude.

Beyond Math Skills: Measuring Quantitative Reasoning in Context

Nathan D. Grawe

In the opening paragraphs of a study sponsored by the National Council on Education and the Disciplines, Lynn Arthur Steen (2001) notes, "Unfortunately, despite years of study and life experience in an environment immersed in data, many educated adults remain functionally innumerate" (p. 1). Steen's call to action has been met by an explosion of quantitative reasoning (QR) programs. (The terms *quantitative reasoning, quantitative literacy,* and *numeracy* are used interchangeably in this literature.) A 2009 Mathematical Association of America survey of U.S. colleges found that two-thirds of institutions host a quantitative support center (see Schield, 2010). Moreover, the survey reports nearly half of institutions have developed quantitative requirements transcending the traditional math/science distribution. Clearly, much innovation has taken place.

Are these efforts effective? The first challenge in answering this question is defining the concept. Through assessing thousands of student papers and reading the theoretical literature, Carleton's Quantitative Inquiry, Reasoning, and Knowledge (QuIRK) initiative has arrived at a working definition: the habit of mind to consider both the power and limitations of quantitative evidence in the evaluation, construction, and communication of arguments in public, professional, and personal life. (See Grawe and Rutz, 2009; Lutsky, 2008; and Rutz and Grawe, 2009, for a full discussion of this definition and its implications for programming.) This definition is broad, intended to be inclusive of a range of QR conceptions. As we speak with researchers and teachers at other institutions, we have come to recognize QR as a multifaceted domain that plays out in different

Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ir.379

ways in different disciplines. A narrow definition inevitably excludes sections of the campus from the discussion—an outcome that would undermine our effort to change campus culture.

This definition contains four clear facets. First, QR requires command of mathematical skills. Students cannot appreciate the strengths and limitations of tools they do not possess. In a campus talk, Deborah Hughes Hallett (2008) suggested the necessary skills: arithmetic, percentages, graphs, estimation, elementary probability and statistics, basic geometry of measurement, and comprehension of basic growth patterns (linear vs. exponential). The absence of calculus, or even algebra, from this list is not to suggest that these tools cannot be employed by the quantitatively literate, but only that these are not essential skills. As Steen (2004) eloquently puts it, QR is “sophisticated reasoning with elementary mathematics rather than elementary reasoning with sophisticated mathematics” (p. 9).

Steen’s quote points to the second facet of QR: it involves application in context (see De Lange, 2003; Steen, 2001, 2004; Richardson and McCallum, 2003; and Bok, 2006). Just as the ability to sound out words or recite grammar rules does not guarantee the comprehension and fluency of literacy, quantitative methods are not completely owned unless the student can apply them to problems in a variety of contexts.

Third, QR involves communication (see Brakke, 2003; De Lange, 2003; and Wiggins, 2003). It may be easy to see that effective communication in the twenty-first century requires facility with QR (see Rutz and Grawe, 2010; and Wolfe, 2010). With increasingly easy access to data over the internet and analysis software from basic graphics to statistical analysis tools, unsupported quantitative claims have become less and less acceptable. Max Frankel (1995) writes in the *New York Times Magazine,* “Deploying numbers skillfully is as important to communication as deploying verbs” (para. 4). Our experience at Carleton shows that even students with strong math skills may come up short in QR from inability to translate that strength into an argument. A reading of the theoretical literature foreshadows this finding.

The final facet of our definition is that QR “describes a habit of mind rather than a set of topics or a list of skills” (Hughes Hallett, 2003, p. 91). We seek to condition our students, on encountering any problem, to ask, “What do the numbers show?” The answer may be that the available quantitative evidence is misleading because it counts only the countable and not the important. Or it may be that QR does not address the question at hand. Recognition of the limitations of quantitative evidence is, after all, a key component of QR. But without what Steen (2001, p. 2) describes as the “predisposition to look at the world through mathematical eyes, to see the benefits (and risks) of thinking quantitatively about commonplace issues, and to approach complex problems with confidence in the value of careful reasoning,” numeracy will be restricted to prompted course work or directed academic exercise.

Given this definition, it might be argued that quantitative and qualitative analyses are merely two alternative reflections of an overarching critical thinking. For instance, just as instructors of numeracy warn their charges to consider the construction of variables, teachers of qualitative approaches caution students to define terms. Similarly, an advocate of QR who demands students eschew "weasel words" such as *many* and *few* in place of actual figures is mimicked by the humanist demanding that pupils demonstrate precision in language. These connections may be true at an abstract level, but QR requires specific skills not usually required by generic critical thinking. For example, QR necessitates an understanding of what statistical significance means and how it is distinct from practical significance or the ability to effectively communicate information through charts and tables. It is exactly this that motivates a growing literature teaching concepts specific to thinking critically with quantitative evidence. (For examples, see Few, 2004; Miller, 2004; and Tufte, 2001.)

Direct Assessments of QR

In 2001, the National Academy of Sciences (NAS) convened a National Forum on Quantitative Literacy. Among the five major findings (summarized in Steen, 2004), the forum concluded, "QL is largely absent from our current systems of assessment and accountability" (p. 9). This is not to say there were no assessment tools. Many national surveys ask students about their comfort with, and self-assessed ability in, the use of quantitative methods. But Chun (2002), who summarizes some of the most commonly used national surveys, notes the questionable reliability of such self-reports. Chun disputes whether students understand items in the way survey constructors intended and, even if they do, whether they can or will provide accurate assessments of their own capacities. Acknowledging a viable role for such indirect assessments, Chun concludes, "If we are interested in understanding what students have learned, we should measure what students have learned" (2002, p. 25).

Wiggins (2003) argues the early lack of assessment tools was predictable, given the contextual nature of the subject and its identity as a habit of mind:

> The implications of contextualized and meaningful assessment in [QR] challenge the very conception of "test" as we understand and employ that term. Test "items" posed under standardized conditions are decontextualized by design. . . . [A]ssessment must be designed to cause questioning (not just "plug and chug" responses to arid prompts); to teach (and not just test) which ideas and performances really matter; and to demonstrate what it means to *do* mathematics [p. 125; emphasis in original].

Although traditional tests of math skills aptly measure students' ability to calculate, they obliterate demands of authentic, ill-structured problems—demands to evaluate and select evidence, choose appropriate methods, interpret results, and ask the "right" question. To contrast the nature of QR with traditional test design, six times Wiggins uses the word *messy* to describe the practice of QR.

Wiggins (2003) is undoubtedly correct in concluding that the abstract and clean environment of traditional standardized math tests cannot capture QR. However, this should not be taken to mean such assessments have no bearing on the subject. If QR requires a basic skill set, then traditional tests of students' abilities to perform calculations or comprehend graphs and tables speak to a critical facet of QR. To a degree, tests such as the SAT math exam or the CAAP mathematics subsection can serve such a purpose. Yet Steen's warning (2004) that QR values sophistication of the application over sophistication of the tool implies that scores on these mathematics exams are only proxies for aptitude in the skills required by QR. For instance, for the purposes of QR skills assessment, trigonometry and geometry play too large a role, while estimation and knowledge of basic magnitudes (such as knowing the world's population is better approximated as six billion than six trillion) is entirely overlooked.

Context-Rich Multiple-Choice Instruments. Since the 2001 NAS forum, several institutions have worked to develop alternative math skills tests that better reflect critical QR skills. Even as he noted the limitations of such tests, Steen (2004) offered sample questions to jump-start the effort. The QR program at Wellesley College (http://www.wellesley.edu/QR/) has created a full exam to identify entering students in need of skills training. Macalester College has developed a test—the Macalester Learning Assessment (MLA)—that includes multiple-choice items testing QR skills and knowledge of basic magnitudes, which they use to assess development of QR in the general education curriculum (see Bressoud, 2009, for sample items).

Of these efforts, perhaps the most developed is James Madison University's Quantitative Reasoning Test (see http://www.madisonassessment.com/view-demo for sample items). The JMU instrument is designed to evaluate student ability in two dimensions: (1) the ability to "use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomenon"; and (2) the capacity to "discriminate between association and causation, and identify the types of evidence used to establish causation" (Sundre, 2008, p. 4). The twenty-five-minute, multiple-choice test has been developed with large samples from JMU, including more than fourteen hundred first-year students in fall 2007 (Sundre, 2008). In that sample, both computer- and paper-based administrations proved "moderately" to "substantially" reliable (Cronbach's alpha = 0.60 and 0.65, respectively), in the terminology of Landis and Koch (1977). A spring 2008 examination of sophomores showed a slightly higher level of reliability among this more advanced student group. JMU has recently collaborated to administer the test at several other institutions, with similar success. Although the test has not yet been given to a large number of institutions, were it widely applied it could yield nationally comparable data. Readers interested in using the test should refer to its description at http://www.jmu.edu/assessment/wm_library/QR.pdf.

Alternatives to Multiple-Choice Tests. Assuming test designers can meet the challenge of incorporating the complexities of context into a multiple-choice framework, the strengths of this approach are obvious. Multiple-choice questions can assess student ability to identify the correct answer to a prescribed calculation. Moreover, students are familiar with such testing environments, scores from computer-administered tests can be calculated instantaneously, competency scores are easily calculated, and administration costs can be quite low. However, Richard Hersh (2010), co-director of the Collegiate Learning Assessment (CLA), questions whether even the best-designed test can capture the richness of complex learning outcomes such as critical thinking or quantitative reasoning. Hersh writes, "Life is not a series of multiple choice questions" (2010, p. 1). In particular, he notes the need to evaluate students' capacity "to use [their] critical thinking and analytical reasoning skills to assemble and reconcile seemingly contradictory information to deal with problems that are not easily defined" (p. 1).

Hersh's argument highlights several aspects of the working definition of QR presented above that are difficult to incorporate into any multiple-choice item. First, students must not only comprehend others' quantitative arguments in text, chart, and table; they must also be able to construct their own. Second, students must be able to discern differences in source quality. Finally, they must navigate the ambiguities of incomplete information and contradictory evidence. In Wiggins's words (2003), they must navigate the messy terrain of ill-structured, context-rich problems. This, after all, is what we do in our own work and in our lives as critical consumers and creators of arguments.

The CLA responds to these assessment challenges with a nationally comparable instrument that includes open-ended prompts with essay response. For the purposes of QR assessment, the most relevant section of the CLA is its "performance task" (see http://www.collegiatelearning assessment.org/ for a more complete discussion of the CLA, including descriptions of the analytic writing sections of the exam). Students are presented with an authentic ill-structured task. For instance, a retired prompt asks students to advise a company on its impending decision to buy an airplane for its sales staff in the wake of a crash. A "document library" offers nearly ten sources of information of varying quality that might be incorporated into brief essays, allowing students to identify and evaluate alternative perspectives. Though not all of the performance tasks involve QR, roughly 80 percent do. In asking students to construct an argument, weigh the credibility of conflicting sources, and grapple with an ill-structured problem with no clear right answer, the CLA represents an interesting counterpoint to QR-skills testing. As a national measure administered at several hundred institutions, it also produces evidence that can be used to compare student performance across campuses.

That said, the CLA's input into QR assessment is oblique at best, for two reasons. First, the creators of the instrument have intentionally taken a holistic approach that incorporates critical thinking, analytical reasoning, written communication, and problem solving. As discussed in the introduction, quantitative reasoning is a subset of this broad domain, making it difficult to interpret the scores relative to QR ability. Second, like all assessment tests, the CLA prompts students to consider particular evidence. The CLA's approach is unquestionably more open-ended than that of a traditional exam, but the presence of quantitative data at the very least prompts students to consider using that evidence in their work. Indeed, weaker students might interpret the presence of the provided data as a requirement that it be incorporated into successful responses. By contrast, authentic prompts offer no guidance as to what content should be included, leaving it up to the individual to decide what evidence should be sought. In prompting consideration of quantitative evidence, the CLA cannot discern students' habit of mind to actively seek out this information.

QR-specific open-ended, standardized prompts have been suggested by several authors. For instance, Steen (2004) and Macalester's MLA propose tests that mix multiple-choice items with brief essay prompts. Similarly, Wiggins (2003) suggests a sample instrument composed of eleven essay prompts. The MLA continues to be developed, and formal evaluation of its reliability and validity remain to be completed.

Carleton's QuIRK initiative has responded to the assessment challenge by stepping away from standardized prompts altogether, adapting writing portfolio analysis for the purpose of QR evaluation. As described in detail in Grawe, Lutsky, and Tassava (2010), QuIRK applies a QR rubric to student work drawn from the college's required sophomore writing portfolio (the rubric and associated codebook can be found at http://serc.carleton.edu/quirk/Assessment/index.html). By looking for evidence in work written for purposes other than QR assessment, we capture student behavior in the absence of a directed task. Through this approach, we can begin to assess students' habits of mind—whether or not they have been conditioned to ask "What do the numbers show?" Furthermore, many of the assignments allow students to choose their own sources. This requires them to weigh alternative pieces of evidence and consider source credibility, both key elements of numeracy in the modern world.

Rubric scorers evaluate the relevance (three categories), extent (three categories), and quality (four-point scale) of QR exhibited in the sampled paper. Extent and quality are probably clear; relevance bears some explanation. Our reading of student papers reveals two categories of QR relevance. The first is central relevance, a case in which quantitative evidence addresses a central question, issue, or theme of the work. For example, in an essay on the deterrent effect of capital punishment, QR is of central relevance.

The second category of QR relevance is peripheral relevance. QR is peripherally relevant to a paper when quantitative evidence would promote useful detail, enrich descriptions, present background, or establish frames of reference. As Miller (2004) writes, "Even for works that are not inherently quantitative, one or two numeric facts can help convey the importance or context of your topic" (p.1). For example, when introducing an argument that the combination of three philosophical definitions of poverty results in a richer conception of the condition than any one perspective alone, a student might choose to generate interest in her argument by comparing the distribution of income in the United States with that in other developed countries. QR is not at the core of this argument, and many QR-irrelevant introductions might be imagined, but the rhetorical strategy chosen by the writer calls out for quantitative evidence.

Given the potential for subjectivity among scorers, interrater reliability is a real concern with instruments of this type. Grawe, Lutsky, and Tassava (2010) report strong reader agreement in the dimensions of QR relevance and extent. Complete agreement was achieved in 75 percent and 82 percent of cases, respectively (corresponding Cohen's kappa = 0.611 and 0.693). The four-point quality scale proved slightly less reliable (perfect agreement in 67 percent of cases; kappa = 0.532), though substantial agreement was achieved when the scale was collapsed to three categories (perfect agreement in 78 percent of cases; kappa = 0.653).

The reliability of an assessment approach such as QuIRK's is undoubtedly contextual. Readers' agreement in evaluation of student work hinges on a common understanding of both the meaning of the instrument and the range of outcomes to be expected by a given student population. This claim is supported by a feasibility study conducted at Morehouse College. In reading Carleton papers, Morehouse faculty reliably applied the rubric to papers written by Morehouse students. Perfect agreement in QR quality was achieved in 82 percent of papers (kappa = 0.74). However, when those same readers scored papers written by Carleton students, reliability was notably lower, with agreement in only 52 percent of cases (kappa = 0.34). If this finding bears out in repeated study, it suggests that the QuIRK rubric can be employed effectively within a campus context but may not produce data that are comparable across institutions. Alternatively, it may mean that readers scoring works from campuses other than their own require greater training and norming to achieve useful levels of agreement.

Assessment at the Classroom, Program, and Institution Levels. Because the instruments discussed above are relatively young, most have not been thoroughly examined in a variety of contexts to formally determine their applicability. However, the details of their construction suggest some obvious limitations. QR skills assessments such as JMU's QR test have promise at all three levels of evaluation. In the case of the JMU test, reliability has been examined at the student level. Presumably, aggregations to a group of students would be more reliable still.

By contrast, the CLA and QuIRK instruments were designed for institution-level assessment. In the case of the CLA, the particular context embedded within a given performance task may not fit any one student, and so prove unreliable. In the case of the QuIRK approach, an assignment may have no QR relevance, or once again the context of an assignment may lead a student to do poorly in that specific situation. Alternatively, although the prompt may elicit QR-relevant work from one student, another student may pursue a direction for which quantitative evidence plays no role. Even in a reasonably large section, only a fraction of students may take up a QR-relevant approach, resulting in a sample size too small to analyze meaningfully. These problems are widely recognized issues pertaining to all performance-based instruments (Chun, 2002). In some sense, this is the price of authentic, highly contextualized performance tasks. For this reason, both instruments call for a large sample of students.

Fostering Improvements in Student Learning

As the citation dates seen above suggest, assessment of QR is too new to generate clear guidance as to "what works" to improve student achievement. We can, however, identify three basic approaches to QR instruction and report preliminary assessment findings. The three curricular strategies differ primarily in their breadth. The most conservative construction casts QR as a context-rich version of mathematics. A second, more expansive implementation of QR sees opportunity for instruction throughout the sciences and in some social sciences. The third and broadest conception embeds QR in courses across the curriculum, including in the traditionally less-quantitative divisions of arts, literature, and humanities.

The mathcentric approach is usually implemented in one of two ways: as "math for the liberal arts" or as statistics. Surely there is much to be gained from these mathcentric courses. Recently developed curricular materials for these types of courses constitute excellent teaching tools (see Madison and others, 2010). However, if disciplines beyond mathematics fail to engage in teaching QR, we risk sending an unintended message that QR is relevant only to mathematics. As Steen (2004) observes, "If [QR] remains the responsibility solely of mathematics departments—especially if it is caged into a single course such as 'Math for Liberal Arts'—students will continue to see it as something that happens only in the mathematics classroom" (p. 18). In addition, it seems unlikely professors in a single discipline can represent the full range of context-rich practice such as QR.

Embedding QR throughout the sciences is a large step toward greater contextual variation. JMU represents one such example. They have combined their QR test with a complementary scientific reasoning (SR) test to assess the effectiveness of the "natural world" cluster of their general education curriculum (assessment reports can be viewed at http://www.jmu.edu/assessment/JMUAssess/GenEdOverview.htm). A description of the requirement indicates that courses within the cluster explore both science and math skills and disciplinary content relevant to the study of the natural world. Not surprisingly, nearly all courses satisfying the requirement are found in the science division. JMU's assessment work reveals that student QR performance increases with the number of credit hours completed in cluster three courses and with high school advanced placement experiences that earned cluster three credits, and that QR capacity is positively related to higher grades in cluster three courses. In total, the JMU experience suggests that one effective way to increase QR skill is to require students to take courses in which professors teach to explicit QR learning outcomes.

If it is good to expose students to QR in multiple departments within a division, is it better to integrate QR throughout the curriculum? Bok (2006, p. 134) argues for this approach:

> Numeracy is not something mastered in a single course. . . . Thus quantitative material needs to permeate the curriculum, not only in the sciences but also in the social sciences and, in appropriate cases, in the humanities.

This strategy has been implemented by the QR programs at Carleton, Hollins University, Macalester College, and Wellesley College, and by the Mathematics Across the Community College Curriculum initiative.[1]

Assessment of the QR-across-the-curriculum approach has been initiated at Carleton. Because the strategy embeds QR within courses in all divisions, credit hours in the sciences serve as a poor proxy for QR exposure. Had Carleton's new QR graduation requirement been in place in the past, we might look for differences in performance by number of QR courses completed, but we will need to wait two years to be able to perform that analysis. In the meantime, longitudinal analysis lends support for the across-the-curriculum approach. Between 2005 and 2007 (just before and just after the work of the initiative began), the fraction of Carleton seniors reporting that their "use of quantitative skills is stronger or much stronger" than at college entry increased from 67 percent to 80 percent. Similarly, the fraction of graduates agreeing that QR "will be very important or essential to my life" increased from 54 percent to 68 percent. Most of this growth is due to increases among majors in the arts, literature, and humanities.

Results from QuIRK's assessment of student papers are direct evidence of success. Among papers for which QR was deemed peripherally relevant, the fraction of students employing QR has increased in five years from 28 percent to 53 percent. Similarly, the fraction of these papers rated of good QR quality rose from 9 percent to 32 percent, while the fraction rated of poor quality fell by an equal amount. A new research project examining the effects of participation in QR first-year seminars has produced promising initial results. The Carleton curricular approach to QR is less concentrated than that at JMU, but the conclusion seems to be the same: intentional teaching with attention to specific learning outcomes can improve students' QR ability.

Conclusions for Future Assessment

Despite the relative youth of QR assessment, several lessons for future assessment seem clear. First, context matters. Because QR itself is grounded in context, our assessment tools must be as well. This presents a challenge for traditional approaches to assessment, which attempt to eliminate contextual aspects of items. Nevertheless, the work at JMU and Macalester suggests it is possible to create context-rich multiple-choice items to assess QR skill competencies.

Second, no matter how carefully designed, multiple-choice exams do not capture several key aspects of the field. Students' ability to construct arguments with quantitative evidence and the habit of mind to seek out that evidence in the first place might be better assessed in essays or oral talks. The experience of the Carleton QuIRK rubric shows it is possible to develop a reliable rubric for that task.

Finally, the multifaceted nature of QR calls for a multifaceted approach to its assessment. Without directed prompts, it is impossible to test a full range of QR skills. For this purpose, the multiple-choice test is both efficient and effective. At the same time, discerning students' habits of mind requires an approach that gives students more freedom. Given this inherent tension, multiple instruments are almost surely necessary. As the educational innovations found in the Mathematical Association of America's survey mature, it will be critical to gain greater clarity as to what works in preparing students to navigate "the twenty-first century . . . world awash in numbers" (Steen, 2001, p. 1).

Note

1. Carleton: see http://serc.carleton.edu/quirk/CarletonResources/curricular_materials.html, or http://www.mac3.amatyc.org/education/education.html. Hollins University: see http://www1.hollins.edu/homepages/hammerpw/qrhomepage.htm#applied. Macalester College: see http://www.macalester.edu/qm4pp/program/index.html#qt. Wellesley College: see http://www.wellesley.edu/QR/overlay.htm. Mathematics Across the Community College Curriculum initiative: see http://www.mac3.amatyc.org/index.htm.

References

Bok, D. *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More.* Princeton, N.J.: Princeton University Press, 2006.

Brakke, D. F. "Addressing Societal and Workforce Needs." In B. L. Madison and L. A. Steen (eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*. Princeton, N.J.: National Council on Education and the Disciplines, 2003.

Bressoud, D. "Establishing the Quantitative Thinking Program at Macalester." *Numeracy*, 2009, *2*(1), Article 3.

Chun, M. "Looking Where the Light Is Better: A Review of the Literature on Assessing Higher Education Quality." *Peer Review*, 2002, *4*(2/3), 16–25.

De Lange, J. "Mathematics for Literacy." In B. L. Madison and L. A. Steen (eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*. Princeton, N.J.: National Council on Education and the Disciplines, 2003.

Few, S. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Oakland, Calif.: Analytics Press, 2004.

Frankel, M. "Word and Image; Innumeracy." *New York Times*, Mar. 5, 1995. Retrieved on June 11, 2010, from http://www.nytimes.com/1995/03/05/magazine/word-image-innumeracy.html?scp=1&sq=Max%20Frankel%20Innumeracy&st=cse.

Grawe, N. D., Lutsky, N. S., and Tassava, C. J. "A Rubric for Assessing Quantitative Reasoning in Written Arguments." *Numeracy*, 2010, *3*(1), Article 3.

Grawe, N. D., and Rutz, C. A. "Integration with Writing Programs: A Strategy for Quantitative Reasoning Program Development." *Numeracy*, 2009, *2*(2), Article 2.

Hersh, R. H. "Life Isn't a Multiple Choice Question." Retrieved on April 27, 2010, from http://www.cae.org/content/pdf/HershLifeIsntAMultipleChoiceQuestion.pdf.

Hughes Hallett, D. "The Role of Mathematics Courses in the Development of Quantitative Literacy." In B. L. Madison and L. A. Steen (eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*. Princeton, N.J.: National Council on Education and the Disciplines, 2003.

Hughes Hallett, D. "How do We Teach Quantitative Reasoning? Foster a Curricular Conspiracy." Presentation at Carleton College's Perlman Center for Teaching and Learning, Apr. 24, 2008.

James Madison University Center for Assessment and Research Studies. "Cluster Three: The Natural World Cluster Assessment Results and Interpretation." 2006. Retrieved May 12, 2010, from http://www.jmu.edu/assessment/JMUAssess/Gened/cluster_3_report_sp06.pdf.

Landis, J. R., and Koch, G. G. "The Measurement of Observer Agreement for Categorical Data." *Biometrics*, 1977, *33*(1), 159–174.

Lutsky, N. S. "Arguing with Numbers: A Rationale and Suggestions for Teaching Quantitative Reasoning Through Argument and Writing." In B. L. Madison and L. A. Steen (eds.), *Calculation vs. Context: Quantitative Literacy and Its Implications for Teacher Education*. Washington, D.C.: Mathematical Association of America, 2008.

Madison, B. L., Boersma, S., Diefenderfer, C. L., and Dingman, S. W. *Case Studies for Quantitative Reasoning: A Casebook of Media Articles*. Boston: Pearson Custom Publishing, 2010.

Miller, J. E. *The Chicago Guide to Writing About Numbers*. Chicago: University of Chicago Press, 2004.

Richardson, R. M., and McCallum, W. G. "The Third R in Literacy." In B. L. Madison and L. A. Steen (eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*. Princeton, N.J.: National Council on Education and the Disciplines, 2003.

Rutz, C. A., and Grawe, N. D. "Pairing WAC and Quantitative Reasoning Through Portfolio Assessment and Faculty Development." *Across the Disciplines*, 2009, *6*(Dec.). Retrieved June 11, 2010, from http://wac.colostate.edu/atd/assessment/rutz_grawe.cfm.

Schield, M. "Quantitative Graduation Requirements at US Four-Year Colleges." 2010. Retrieved on June 11, 2010, from http://statlit.org/pdf/2010SchieldJMM.pdf.

Steen, L. A. "The Case for Quantitative Literacy." In L. A. Steen (ed.), *Mathematics and Democracy: The Case for Quantitative Literacy*. Princeton, N.J.: National Council on Education and the Disciplines, 2001.

Steen, L. A. *Achieving Quantitative Literacy: An Urgent Challenge for Higher Education*. Washington, D.C.: Mathematical Association of America, 2004.

Sundre, D. L. *The Quantitative Reasoning Test*, Version 9 (QR-9), 2008. Retrieved on May 10, 2010, from http://www.madisonassessment.com/uploads/qr-9_manual_2008.pdf.

Tufte, E. R. *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press, 2001.

Wiggins, G. "'Get Real!': Assessing for Quantitative Literacy." In B. L. Madison and L. A. Steen (eds.), *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*. Princeton, N.J.: National Council on Education and the Disciplines, 2003.

Wolfe, J. "Rhetorical Numbers: A Case for Quantitative Writing in the Composition Classroom." *College Composition and Communication*, 2010, *61*(3): 452–475.

NATHAN D. GRAWE is associate professor of economics and associate dean of the college and director of the Quantitative Inquiry, Reasoning, and Knowledge (QuIRK) initiative at Carleton College.