2010

A Rubric for Assessing Quantitative Reasoning in Written Arguments

Nathan D. Grawe
Department of Economics, Carleton College, Northfield MN, ngrawe@carleton.edu
Neil S. Lutsky
Department of Psychology, Carleton College, Northfield MN
Christopher J. Tassava
Office of Corporate and Foundation Relations, Carleton College, Northfield MN

A Rubric for Assessing Quantitative Reasoning in Written Arguments

Abstract

This paper introduces a rubric for assessing QR in student papers and analyzes the inter-rater reliability of the instrument based on a reading session involving 11 participants. Despite the disciplinary diversity of the group (which included a faculty member from the arts and literature, two staff members, and representatives from five natural and social science departments), the rubric produced reliable measures of QR use and proficiency in a sample of student papers. Readers agreed on the relevance and extent of QR in 75.0 and 81.9 percent of cases respectively (corresponding to Cohen's $\kappa = 0.611$ and 0.693). A four-category measure of quality produced slightly less agreement (66.7 percent, $\kappa = 0.532$). Collapsing the index into a 3-point scale raises the inter-rater agreement to 77.8 percent ($\kappa = 0.653$). The substantial agreement attained by this rubric suggests that it is possible to construct a reliable instrument for the assessment of QR in student arguments.

Introduction

At its 2001 Forum on Quantitative Literacy, the National Council on Education and the Disciplines concluded, "Quantitative literacy is largely absent from our current systems of assessment and accountability" (Steen 2001). Since the writing of that report, researchers have been busy attempting to fill the gap. However, the very nature of quantitative reasoning (QR) presents a hurdle. Many authors argue that QR involves implementation in context (Bok 2006, 129; De Lange 2003, 80; Richardson and McCallum 2003, 100−102; Steen 2004, 9−10). This is in keeping with the goals of educational initiatives that seek to strengthen students' willingness to use QR in a wide variety of appropriate circumstances and to do so effectively. As Steen writes, "The test of numeracy, as of any literacy, is whether a person naturally uses appropriate skills in many different contexts" (2001, 6).

Taylor (2009) provides a brief survey of current QR assessment efforts. Traditional testing methods use multiple-choice questions or calculation problems to determine whether students have gained basic quantitative skills and understandings. This approach provides test takers with problems that explicitly call upon knowledge of quantitative concepts and tools. Thus standardized assessment of this sort can tell us whether students have the capacity to apply QR knowledgeably when prompted to do so, an important foundational skill for QR; the tests don't, however, show whether students have strengthened a tendency to use that capacity or have developed the skills necessary to deploy the capacity effectively in contexts other than those in the test. It is possible to engineer a standardized test to represent quantitative skills useful or necessary in selected contextual domains (e.g., for scientific reasoning or understanding medical information), but, as Wallace et al. (2009) note, it is important to recognize that demonstrating a skill in the context of a specific test doesn't mean the skill will be generalized to other contexts or will indicate the presence of other skills necessary to employ QR successfully in those other contexts.

Recent authors have also noted that QR extends beyond calculation into the realm of argumentation. For instance, De Lange (2003, 77) and Brakke (2003, 168) emphasize the *communication* of quantitative analysis, presumably including visual presentation through tables and figures in addition to integrating numbers into prose. Others have amplified this idea, framing QR in the context of argument (e.g., Grawe and Rutz 2009; Lutsky 2008; Schield 2008; Steen 200; Taylor 2008). The BBC radio program *More or Less* pithily summarizes this point: "Numbers [are] the principal language of public argument." Our reading of this literature leads us to an understanding of QR that might be summarized as *the habit of mind to consider the power and limitations of quantitative evidence in the*

evaluation, construction, and communication of arguments in personal, professional, and public life.

If QR is meaningful in the context of evaluating and articulating arguments, then it might be useful to develop an assessment method that closely matches our educational goal. We see two possible benefits to this approach. First, it seems plausible that students who prove quite capable in skills-based assessments may not have developed the habit of mind or flexibility to apply those competencies in the context of arguments. Thus, a direct assessment of the use of QR in written argument may prove a more valid measure.[1] Second, the assessment of actual student work can be a powerful formative assessment experience. Confronting faculty directly with what students are or are not doing with regard to quantitative evidence can motivate and guide professional development activity.

Steen (2004, 16) argues that "[QR] requires creativity in assessment, since neither course grades nor test scores provide a reliable surrogate." The rubric for assessing QR in student writing which we propose in this paper is an attempt to answer Steen's call. In the next section we describe the scoring rubric which we have employed in evaluating QR use and proficiency in papers submitted by students for Carleton College's sophomore writing portfolio. The subsequent two sections describe our methods for testing the reliability of the instrument and give results. We conclude with a discussion of the power of applying the rubric as a formative assessment tool and directions for possible future research applications.

A Rubric for Assessing QR in Written Argument

Context for Use

We contend that it is possible to create a reliable instrument for measuring QR in written arguments. The rubric presented here was developed over four years in the context of Carleton's QR initiative. To foster adaptations of our method to match other institutions' goals, an appendix notes some lessons we learned in the development process. The rubric presented in this section is designed to be applied to a sample of student writing to assess QR at an institutional level. In particular, the rubric is *not* designed to evaluate individual students. The papers we assess were not submitted by students for the purpose of showing QR proficiency and frequently, in fact, contain no evidence of QR proficiency one way or the other. Rather, we hope to examine uses of QR as a whole in order to

---

[1] At this time, there is no widely agreed upon measure of "QR aptitude" and so it is impossible to test this hypothesis. One possible avenue for future research would be to analyze the correlations between alternative QR and critical thinking assessment tools. While this would not resolve questions of validity for any of the tools, it might help us to understand better the various facets of QR and how the alternative assessment tools relate to one another.

gain insight into how we can improve instruction at the institution and to compare QR activity between large groups (e.g., the class of 2005 vs. the class of 2010, or students who major in the social sciences vs. those who major in the humanities) in order to discern effects of institution-level programs and curricular reforms.

While the use of quantitative evidence varies by discipline, the rubric presented here guides scorers to assess the degree to which the use/misuse of QR forwards or fails to forward the argument without regard for the department for which the paper was written. This statement may seem counterintuitive given that we have argued above for the importance of context. We would note, however, that the direction is to ignore only one narrow aspect of the context: the identity of the professor who first read the paper. The entire context inherently related to the argument itself remains.

We have two reasons for asking readers to ignore the identity and disciplinary affiliation of the original professor. First, we do not want readers to attempt to insert themselves into the "mind" of another person. It seems likely that our stereotypical understandings of other disciplines are inaccurate and vary from person to person. The result would likely be increased noise in assessment scores. Second, our purpose is to learn how well our institution prepares students to address problems and arguments in their everyday lives. This general education goal transcends disciplinary norms. We believe that we can arrive at agreed upon standards for the use of evidence in this general education context.

Rubric Items

The first section of the scoring sheet asks for identification numbers of both the student and the reader. The scoring sheet is reproduced in Figure 1. The complete codebook which accompanies the scoring sheet can be found on our program Web site.[2]

Next, readers are asked to assess the potential contribution of quantitative information to the paper based on the stated and implied goals of the paper itself (section II of the scoring sheet). In making this determination, scorers are prompted to consider how a reasonable person would view the relevance of QR to the topic chosen by the student. Note that the question is not whether the student actually uses numeric evidence but rather whether the student has chosen a topic for which such evidence would be deemed useful in a strong paper on that topic. Similarly, we are not interested here in the nature of the assignment (though this will be assessed later in the rubric). It is quite conceivable that one student may react to a paper prompt—for instance, on a critical public-policy issue such as

---

[2] Carleton's Quantitative Inquiry, Reasoning and Knowledge (QuIRK) Initiative. http://serc.carleton.edu/quirk (accessed Dec. 4, 2009).

capital punishment—with an argument that cries out for quantitative analysis while another student chooses a response involving no QR aspect.

Quantitative Reasoning in Student Writing Rating Sheet

I. Identification.   Student I.D. #: _____   Reader I.D. #: _____

II. Is QR potentially relevant to this paper?   [rate potential contents of paper, not the assignment]

___ **NO or incidentally only**   ___ **YES, but peripherally only**   ___ **YES, centrally**

III. What is the extent of numerical evidence and quantitative reasoning present in the paper?
[See: "Employs QR Criteria";   **Note:** This is *not* a rating of the quality of the QR shown, only its presence.]

___ **rating of 1-3,** review attached criteria

IV. OVERALL ASSESSMENT of Quality of implementation, interpretation, and communication of QR:

___ **rating of 1-4,** review attached criteria

V. Problematic characteristics of the QR present in the paper: [check all issues that detract significantly from the reader's understanding of the information presented.]

___ **Uses** ambiguous words rather than numbers.

___ **Fails** to provide numbers that would contextualize the argument.

___ **Fails** to describe own or others' data collection methods.

___ **Doesn't** evaluate source or methods credibility and limitations.

___ **Inadequate** scholarship on the origins of quantitative information cited.

___ **Makes** an unsupported claim about the causal meaning of findings.

___ **Presents numbers without comparisons that might give them meaning.**

___ **Presents** numbers but doesn't weave them into a coherent argument.

VI. Does the assignment explicitly call for the use of QR in the paper?

___ **YES**   ___ **NO**   ___ **NO ASSIGNMENT PRESENT**

Figure 1. Scoring rubric.

The rubric allows three possible responses: No relevance, peripheral relevance, and central relevance. Examples of papers which likely fall in the first category might include an examination of the role of Confucianism in the downfall of the Han dynasty or a comparison of the depictions of Lucretia in paintings by Rembrandt and Gentileschi.

Our past reading of student work suggests that papers for which QR is relevant can actually involve quantitative evidence in either a central or a

peripheral way. Papers for which QR is centrally relevant—in which numbers address a central question, issue, or theme—are perhaps the most obvious "QR papers." What, if any, are the deterrent effects of capital punishment on crime? How does the genetic frequency in two populations of insect larvae inform our understanding of the processes that lead to heterogeneity across populations?

But, as Jane Miller (2004, 1) notes, "Even for works that are not inherently quantitative, one or two numeric facts can help convey the importance or context of your topic." This peripheral use of QR employs numbers to provide useful detail, enrich descriptions, present background, or establish frames of reference. For instance, a paper tracing possible psychogenic pain mechanisms is centrally focused on the neuroscience of physical sensation. But a strong paper on this topic might use numbers to describe the incidence of the phenomena in an introductory paragraph. Similarly, a student might open an observational essay evaluating the nature of community in a contemporary American mall by discussing the demographics of mall shoppers or the geographic distribution of malls. Such a paper would be immeasurably improved by the use of precise quantitative information rather than unsubstantiated claims that "many" people "often" do such and such.

After assessing QR relevance, readers evaluate the extent of quantitative evidence present in the paper (section III of the scoring sheet) by choosing one three possible categories:

1. No explicit numerical evidence or quantitative reasoning. May include quasi-numeric references (i.e., "many," "few," "most," "increased," "fell," etc.).
2. One or two instances of explicit numerical evidence or quantitative reasoning (perhaps in the introduction to set the context), but no more.
3. Explicit numerical evidence or quantitative reasoning is used throughout the paper.

At one extreme, the paper might include no explicit numerical evidence or quantitative reasoning. At the other, explicit QR might be present throughout. In between, an essay might include one or two instances of explicit QR (most often seen in an introduction or conclusion, though sometimes present in a single example or element of the argument). At this point, scorers are not asked to consider the *quality* of the evidence presented—which may be brilliant or wholly fallacious. Rather, scorers are asked to gauge the degree to which students call upon quantitative evidence in support of their arguments.

Sections IV and V of the scoring sheet call for quality assessment. Because it makes little sense to evaluate the use of QR when QR is irrelevant to the paper, these sections are not scored for QR-irrelevant essays. In section IV the reader records an evaluation of the overall quality of the use of QR in the paper on a

scale of 1 to 4. In high-scoring papers, the use of QR enhances the effectiveness of the paper, advancing the argument. By contrast, in low-scoring papers, the ineffectiveness or absence of QR substantially weakens the argument.

Table 1. Rubric Language for Assessing Quality of QR

A. In Papers where QR is Centrally Relevant

Quality Score			
1	**2**	**3**	**4**
Use of numerical evidence is so poor that either it is impossible to evaluate the argument with the information presented or the argument is clearly fallacious. Perhaps key aspects of data collection methods are missing or critical aspects of data source credibility are left unexplored. The argument may exhibit glaring misinterpretation (for instance, deep confusion of correlation and causation). Numbers may be presented, but are not woven into the argument.	**The use of numerical evidence is sufficient to allow the reader to follow the argument. But there may be times when information is missing or misused.** Perhaps the use of numerical evidence itself is uneven. Or the data are presented effectively, but a lack of discussion of source credibility or methods makes a full evaluation of the argument impossible. Misinterpretations such as the confusion of correlation and causation may appear, but not in a way that fundamentally undermines the entire argument.	**The use of numerical evidence is good throughout the argument.** Only occasionally (and never in a manner that substantially undermines the credibility of the argument) does the paper fail to explore source credibility or explain methods when needed. While there may be small, nuanced errors in the interpretation, the use of numerical evidence is generally sound. **However, the paper may not explore all possible aspects of that evidence.**	**The use of numerical evidence is consistently of the highest quality.** When appropriate, source credibility is fully explored and methods are completely explained. Interpretation of the numerical evidence is complete, considering all available information. There are no errors such as confusion of correlation and causation. **This paper would be an excellent choice as an example of effective central QR to be shared with students and faculty.**

B. In Papers where QR is Peripherally Relevant

1	**2**	**3**	**4**
Fails to use any explicit numerical evidence to provide context. The paper is weaker as a result. **This paper shows no attempt to employ peripheral QR.**	Uses numerical evidence to provide context in some places, but not in others. The missing context weakens the overall paper. Or the paper may consistently provide data to frame the argument, but fail to put that data in context by citing other numbers for comparison. **Ultimately, the attempt at peripheral use of QR does not achieve its goal.**	The paper consistently provides numerical evidence to contextualize the argument when appropriate. Moreover, numbers are presented with comparisons (when needed) to give them meaning. However, there may be times when a better number could have been chosen or more could have been done with a given figure. **In total, the peripheral use of QR effectively frames or motivates the argument.**	Throughout the paper, numerical evidence is used to frame the argument in an insightful and effective way. When needed, comparisons are provided to put numbers in context. **This paper would be an excellent choice as an example of effective peripheral QR to be shared with students and faculty.**

Because expectations for QR differ by whether the use (or missed use) was central or only peripheral to the argument, we provide distinct scoring language

for each category (Table 1). Table 1a presents guidelines for centrally relevant papers. The key feature of a paper given a score of 1 is that the use (or absence) of QR is so problematic that the argument fundamentally fails: either it is impossible to evaluate the argument given the provided evidence or the argument is clearly fallacious. If the use of QR does not entirely undermine the argument and yet important quantitative information is missing or misused, the paper is given a quality score of 2. Only if the use of numerical evidence is sound throughout the paper does it receive a 3 or 4. Readers give a score of 4 if they view the paper as exemplary in the quality, insightfulness, and completeness of QR implementation.

The scoring language for peripheral papers (Table 1b) is necessarily different because the use of QR in a peripheral context is only to frame a discussion—it is not the crux of the argument. Despite these differences, the scoring logic is very similar. A score of 1 denotes a paper that fails to provide explicit numerical context and so is weaker as a result. Just as in the case of a centrally relevant paper, this score indicates that the use or missed use of QR undermines the paper's purpose. A score of 2 indicates that the student did employ QR but in an uneven way or such that the peripheral use does not achieve its goal. Once more, a 3 means the paper is consistently successful in its uses of QR to set the context or frame the argument, and again a 4 denotes an exemplar of peripheral QR use.[3]

Repeated reading also highlighted several problematic characteristics common to first-year and sophomore papers. In section V of the scoring sheet, scorers code for whether the presence of the following eight problems detracts significantly from the reader's understanding of the information presented (the figures in parenthesis indicate the frequency that each issue was observed in the scoring session described in the next section of this paper):

- Uses ambiguous words rather than numbers (27.1%).
- Fails to provide numbers that would contextualize the argument (31.9%).
- Fails to describe own or others' data collection methods (6.9%).
- Doesn't evaluate source or methods credibility and limitations (11.1%)
- Inadequate scholarship on the origins of quantitative information cited (7.6%).
- Makes an unsupported claim about the causal meaning of findings (11.8%).
- Presents numbers without comparisons that might give them meaning (15.3%).
- Presents numbers but doesn't weave them into a coherent argument

---

[3] In four years of paper reading, our group has repeatedly encountered a number of paper types. The online codebook lists a number of these along with the typical scores such papers would receive. We review this as part of the norming session before scoring. The codebook can be found at http://serc.carleton.edu/files/quirk/quirk_rubric.v5.doc (accessed Dec. 4, 2009).

(12.5%).

In this section of the rubric readers are scoring for the *presence* of a problem. For instance, if a student does a nice job distinguishing correlation from causation in one section of the paper and then glaringly fails to do so in a subsequent section, then we code the paper as exhibiting this problem.

Finally, section VI asks raters to read the assignment (if the assignment prompt was submitted with the paper) to determine whether it explicitly calls for the use of QR. This information will be useful in the future as we examine student choices in the "real world" context of problems that do not explicitly prompt quantitative analysis. This item was placed at the end of the scoring sheet to reduce the chance that readers would consider the department from which the paper was written when making quality assessments.

Methods for Evaluating Instrument Reliability

The readers in our assessment responded to a request posted to an email list of faculty and staff who had expressed interest in Carleton's QR initiative. The 11 participants represented a diverse group including:

- 9 faculty and 2 staff.
- 3 full professors, 4 associate professors, 1 un-tenured assistant professor, and 1 lecturer.
- 3 natural scientists (from 2 departments), 5 social scientists (from 3 departments), and 1 faculty member from a department in the arts and literature division.
- 4 men and 7 women.
- 2 individuals who had not read portfolios using a QR rubric prior to this event.

Participants were paid $150 for the four-hour reading session.

We applied the rubric to a sample of papers submitted by students of the class of 2010 as part of the College's writing portfolio. Collected from students at the end of the sophomore year, Carleton's portfolio must include three to five essays written in at least two of the four college divisions and demonstrate competency in five areas: observation, analysis, interpretation, documented sources, and thesis-driven argument. Copies of associated assignments are requested, but many students fail to include them. Students also submit a reflective essay explaining how the portfolio represents their writing. Carleton currently has no QR graduation requirement. Students are required to complete three courses in mathematics or natural sciences. Many complete these requirements by the end of the sophomore year, but they are not required to do so.

We excluded from our sample all of those portfolios which initially received less than a passing mark when assessed by the Writing Program (approximately 5% of all portfolios). Following the guidance of Carleton's Institutional Review Board, we also excluded portfolios from the roughly 15% of students who chose not to allow their work to be used for research purposes.

From the resulting population, we drew a random 50% sample of portfolios (207 in total). From each of these portfolios, we randomly chose one of the papers submitted by the student to fulfill the categories of analysis, interpretation, or observation.[4]

The assessment session began with a norming activity. First, we read through the rubric and its codebook, discussing any questions readers had. Then each reader was asked to score a common set of three papers. In between each scoring, the group discussed its ratings and talked about variation among raters to settle any misunderstandings.[5]

Table 2
Summary Statistics Describing Students who Wrote Scored
Papers and Courses for which the Papers Were Written

	Percent		Percent
Student demographics			
Sex		*Ethnicity*	
Male	43.5	White	82.6
Female	56.5	African American	7.3
		Hispanic	4.4
		Asian	4.4
		No response	1.5
Course characteristics			
Division		*Level*	
Lower	67.7	Arts and Literature	33.8
Middle	30.8	Humanities	17.7
Upper	1.5	Natural Sciences	23.5
		Social Sciences	20.6
		Interdisciplinary	4.4

---

[4] Our intention in selecting papers from these three categories was to increase the likelihood of encountering QR-relevant papers. Because the instructions given to students with the writing portfolio explicitly mention data in descriptions of these three categories, we suspected students would be more likely to submit data-related papers under these headings. Subsequent study of the course of origin of papers submitted to the portfolio suggests that students may be submitting many QR-rich papers under the documented sources category. In the future, we intend to draw randomly from all submitted papers.

[5] Intentionally, the three papers included both strong and weak examples and papers that were both peripherally and centrally QR-relevant.

After the norming work, readers began scoring papers from the sample, which was arranged alphabetically by the student's last name. Each paper was read by two readers. Readers were not matched. At the end of the allotted time, the group had read and scored papers from 72 students. Table 2 presents summary statistics describing the students who wrote scored papers and the courses for which the papers were written. The summary statistics confirm the representative nature of the sample with demographics more or less matching the College as a whole. The relative over-representation of arts and literature courses likely reflects the fact that these courses yield more paper assignments and so are more likely to show up in the writing portfolio. Nevertheless, all four divisions are well represented in the sample. The distribution over lower-, middle-, and upper-level courses shows a large quantity of introductory coursework and less upper-level work, as expected given the timing of portfolio collection at the end of the students' sophomore year.

Results

Table 3 presents results concerning the potential relevance and actual extent of QR in students' papers. The table includes the full two-way table of scores by both readers with the percent of total observations in a given cell provided in parenthesis.

We summarize the tables using two measures of inter-rater agreement: the percent of papers scored identically and Cohen's κ statistic (Cohen 1960). The former is simply the sum of the percentages on the main diagonal of the two-way table. The latter corrects this percentage agreement that we would expect readers to achieve by random chance. For instance, if readers randomly assigned scores on an n-point scale according to the uniform distribution, we would expect random agreement in $1/n$ percent of cases. If readers randomly assign scores according to a non-uniform distribution, the probability of agreement is given by

$$\sum_{i=1}^{n} p_i^2$$

where p_i is the fraction of items scored as category i. Cohen's κ statistic reports the degree to which the observed agreement exceeds the expected agreement, relative to the agreement not explained by chance:

$$\kappa = \frac{Actual\ agreement - Chance\ agreement}{1 - Chance\ agreement}.$$

Complete agreement and chance agreement correspond to a κ statistics of 1 and 0 respectively.

The rubric proved quite reliable in assessing QR relevance—the potential contribution of QR to the stated and implied goals of the paper (section II of the rubric). Readers achieved exact agreement in more than three-fourths of cases (Table 3, upper panel). The κ statistic of 0.611 rises to the "substantial" level ($0.6 \leq \kappa \leq 0.8$) defined by Landis and Koch (1977). Only in one case did a reader view QR as centrally relevant while the other saw no relevance.

Table 3
Inter-Rater Reliability of QR Use:
Relevance and Extent

Is QR potentially relevant to this paper?

	No	Peripherally	Centrally
No	26 (36.1)	5 (6.9)	0 (0.0)
Peripherally	4 (5.6)	7 (9.7)	4 (5.6)
Centrally	1 (1.4)	4 (5.6)	21 (29.2)

Percent agreement = 75.0
Cohen's κ = 0.611   Standard error of κ = 0.085

What is the extent of numerical evidence and quantitative reasoning present?

	Score 1	Score 2	Score 3
Score 1	33 (50.0)	2 (2.8)	0 (0.0)
Score 2	7 (9.7)	6 (8.3)	3 (4.2)
Score 3	0 (0.0)	1 (1.4)	17 (23.6)

Percent agreement = 81.9
Cohen's κ = 0.693   Standard error of κ = 0.086

Note: Rubric language for coding extent of QR:
1: No explicit numerical evidence or quantitative reasoning. May include quasi-numeric references (i.e., "many," "few," "most," "increased," "fell," etc.).
2: One or two instances of explicit numerical evidence or quantitative reasoning (perhaps in the introduction to set the context), but no more.
3: Explicit numerical evidence or quantitative reasoning is used throughout the paper.

Agreement about the extent of QR in the papers (section III of the rubric) was even greater (Table 3, lower panel). Exact agreement was achieved in more than

80% of cases ($\kappa = 0.705$).[6] In no case did readers disagree in the extreme with one reader seeing no QR present while the other reported QR throughout the paper.

Comparing the patterns of agreement seen in relevance and extent, we see that in both cases the disagreements are more likely to involve the "middle" categories of "peripheral relevance" and "some QR." In fact, in only one case did one rater score a paper as QR-irrelevant while the other saw it as centrally relevant, and in no case did one rater code the extent of QR as extensive while the other reported no QR. In part, this pattern is predictable because the highest and lowest categories are adjacent to only one other category while the middle rating has potential for disagreement on both the high and low end.

However predictable the pattern, it raises real concerns. For instance, Grawe and Marfleet (2009) report that QR is relevant to over half of papers submitted to Carleton's writing portfolio and, of particular note, quantitative relevance has a role in all divisions of the college. Even among papers written for courses in art, literature, and humanities, rubric scorers deemed QR relevant over one-third of the time. Not surprisingly, peripherally relevant papers make up a large portion (73%) of potentially QR-relevant work in these "traditionally non-quantitative" fields. If peripherally relevant papers provide an important opportunity to expand QR across the curriculum, it would be nice to see greater inter-rater agreement on these papers.

For those wishing to adopt the rubric for applications requiring greater agreement in the middle categories of relevance and extent, we would suggest a revised assessment protocol which required resolution of disagreements. This might be done by having the two raters negotiate their differences, or the paper could be given to a third reader to break the tie.

Agreement in evaluations of QR quality (section IV of the rubric) was somewhat lower (Table 4). This result is not surprising; disagreements concerning QR relevance easily leads to disagreements over quality due to the different rubric language depending on the category of relevance. The upper panel of Table 4 shows that readers nevertheless achieved exact agreement in over 65 percent of all cases ($\kappa = 0.532$).[7] This level of reliability lies in the "moderate" range using the terminology of Landis and Koch ($0.4 \leq \kappa \leq 0.6$). Examining the two-way table, readers more reliability differentiated papers of exceptionally low and exceptionally high quality. The lower panel of Table 4 shows that reliability

---

[6] In ten cases, readers failed to code the extent of QR. In eight of the ten, the second reader coded extent as none or incidental. The most likely explanation for the missing coding is that the reader found no QR. Assuming this explanation, we recoded these ten missing cases as showing no QR.

[7] Three readers gave QR quality assessments in 16 cases in which they determined QR to be irrelevant to the paper. Because it is difficult to understand how QR could be present if irrelevant or assessed if not present, these quality assessments were recoded as "no score." The results are not substantially altered if the scores are left unchanged.

improves when the scores are collapsed into a three-category scale by combining the middle two levels (scores 2 and 3, according with the two middling quality scores). Using this modified categorization, readers achieved exact agreement in more than 75% of all cases and "substantial" reliability ($\kappa = 0.653$).[8] (Of course, the greater reliability comes with a loss to variation within the data.) These results suggest that the assessment rubric presented in the previous section can be reliably applied in studies of student arguments.

Table 4
Inter-Rater Reliability of QR Quality Using 4- and 3-Category Scales

Overall assessment of quality of QR (4-category quality category):

	No score	1 (poor)	2 (adequate)	3 (good)	4 (exemplary)
No score	27 (37.5)	4 (5.6)	0 (0.0)	0 (0.0)	0 (0.0)
1 (poor)	3 (4.1)	6 (8.3)	0 (0.0)	1 (1.4)	0 (0.0)
2 (adequate)	1 (1.4)	1 (1.4)	6 (8.3)	5 (6.9)	1 (1.4)
3 (good)	1 (1.4)	1 (1.4)	3 (4.2)	9 (12.5)	2 (2.8)
4 (exemplary)	0 (0.0)	0 (0.0)	1 (1.4)	0 (0.0)	0 (0.0)

Percent agreement = 66.7
Cohen's $\kappa = 0.532$ Standard error of $\kappa = 0.068$

Overall assessment of quality of QR (3-category scale):

	No score	1 (poor)	2 or 3 (adequate/good)	4 (exemplary)
No score	27 (37.5)	4 (5.6)	0 (0.0)	0 (0.0)
1 (poor)	3 (4.2)	6 (8.3)	1 (1.4)	0 (0.0)
2 or 3 (adequate/good)	2 (2.8)	2 (2.8)	23 (31.9)	3 (4.2)
4 (exemplary)	0 (0.0)	0 (0.0)	1 (1.4)	0 (0.0)

Percent agreement = 77.8

---

[8] By comparison, the SAT writing exam scores student essays on a 6-point scale. Each essay is read by two readers. Exact agreement is reached in 56% of cases and readers come within one point of each other in another 40% of cases (Camara and Schmidt 2006). While it is impossible to compare perfectly the two rubrics, we might think of collapsing the SAT scale from six categories into three. To a first approximation we might expect that half of the ratings falling within one point of each other would be reconciled in the new three-point scale. Thus, a first approximation of the SAT essay exam's agreement on a three-point scale would be 76% (i.e. 56% + 20%)—the same as achieved here. Given the extensive norming completed by SAT raters—readers must score up to 50 essays before they evaluate actual exams (College Board 2003)—we view this comparison favorably.

Cohen's $\kappa = 0.653$ Standard error of $\kappa = 0.083$

Note: The characteristics "poor," "adequate," "good" and "exemplary" were intentionally not connected with the four quality categories in the scoring rubric because several raters found them distracting. They are attached here for expository reasons only. See previous section for language in the rubric which describes the quality associated with each score.

While the holistic assessment of quality achieved "substantial" reliability, scorers' assessments of *particular* QR problems were more divergent. Table 5 presents the percentage of exact agreement and κ statistics for the eight problem characteristics identified on the rubric. Readers who deemed a paper QR-irrelevant would not score these items, so there are three possible outcomes—problem present, problem not present, and no score given. Readers agreed in approximately two-thirds to three-quarters of cases and achieved "moderate" reliability (κ between 0.429 and 0.532) with but one exception: item "Fails to provide numbers that would contextualize the argument" saw agreement only around half of the time and "fair" agreement ($\kappa = 0.332$). This degree of reliability seems high enough for use in future research but suggests measurement error issues will pose problems of low precision and attenuation bias. Future adaptations of the rubric may be needed before these items can be used as fruitfully as the holistic quality assessment.

Table 5

Inter-Rater Reliability of Problematic Characteristics

	Percent agreement	Cohen's κ	Standard error of κ
Uses ambiguous words rather than numbers.	66.7	0.501	0.080
Fails to provide numbers that would contextualize the argument.	55.6	0.332	0.083
Fails to describe own or others' data collection methods.	73.6	0.489	0.098
Doesn't evaluate source or methods' credibility and limitations.	68.2	0.429	0.092
Inadequate scholarship on the origins of quantitative information cited.	75.0	0.523	0.097
Makes an unsupported claim about the causal meaning of findings.	69.4	0.460	0.091
Presents numbers without comparisons that might give them meaning.	68.1	0.462	0.089
Presents numbers but doesn't weave them into a coherent argument.	70.8	0.489	0.091

Assessment expert Grant Wiggins (2001) writes, "As in book literacy, evidence of students' ability to play the messy game of the [QR] discipline

depends on seeing whether they can handle tasks without specific cues, prompts, or simplifying scaffolds from the teacher-coach or test designer." Unlike traditional QR assessments, student papers provide evidence of student behaviors in the open-ended environment described by Wiggins. When coding assignments (section VI of the rubric), readers achieved exact agreement in almost 90% of all cases ($\kappa = 0.770$).[9] If we exclude the nearly one-half of cases in which the assignment was missing, we find nearly identical results.

The statistics presented in Tables 3−5 suggest that the rubric presented above is reliable in the context of Carleton readers. Our hope is that this approach will be useful for others as well. One way to explore the adaptability of the tool to diverse raters is to examine individual readers' scores relative to the group. If the rubric is robust to broad application, then we would not expect to see significant outliers within our group.

Chi-square goodness of fit tests for equality between each individual's scoring distribution and that of the group as a whole suggest that the rubric is applied similarly by all of the readers.[10] There is little to no evidence that any of the readers produced score distributions that differed substantively from the group as a whole. With 11 readers, each examined on scoring in three dimensions (relevance, extent, and quality) we performed 33 chi-square goodness of fit tests. None had p values of less than 0.05. In practical terms, the reliability statistics reported above are not driven by any one reader. The κ statistics which resulted when individuals are removed one by one are not substantially different from that obtained by the group as a whole. But for one reader's scoring of QR relevance, the changes in κ are all less than 0.1. Of particular importance, with the exception of the same reader, no single reader shifted the reliability of QR quality by more than 0.05. It is worth noting that excluding this reader would have improved the reliability of quality assessment into the "substantial" range even when using a 4-point quality scale.[11]

Because only three of our 11 readers came from outside the natural and social sciences we cannot draw precise predictions about the reliability of the rubric within this group. However, the results above are consistent with the hypothesis that a group of readers drawn from across all divisions of the academy can be trained to apply the rubric reliably.

---

[9] In two cases, individuals failed to score the assignment item. We assume the scorer did not score the assignment because they did not see one present and so recoded these two cases as "no assignment."

[10] A detailed table showing results obtained by removing each reader in turn is available from the authors on request.

[11] The reader in question happens to be one of the participants who had no prior experience scoring student essays for QR proficiency. The one other first-time reader did not affect any of the reliability measures to a substantial degree.

In all of the above, we present reliability of scoring under the assumption that the paper will be read by a single reader. One common way to boost reliability is to require a third reader in cases in which the first two readers disagreed. While we have not completed that exercise, a team at the College of New Jersey is applying this rubric with this three-reader strategy.

Conclusion

This paper presents a rubric for assessing quantitative reasoning (QR) in the context of student-written arguments. In the process of its development, we have found it to be an effective formative assessment tool in at least three senses. First, the process of collectively reading papers through the lens of the rubric has nurtured a focused discussion around the definition of QR, evidence of its presence, assignments that support its development, and professional development activities that might enhance QR instruction. As Grawe and Rutz (2009) describe in detail, these conversations were critical in developing a campus conversation engaging roughly two-thirds of the faculty and ultimately resulted in a new QR graduation requirement. Second, application of the rubric to student work has helped to identify examples of weak and strong student use of QR—examples which have strengthened presentations given to a wide audience at workshops, learning and teaching center seminars, and faculty retreats. Finally, the findings of our assessment work have shaped our programming. For example, recognizing the large fraction of papers for which QR is peripherally relevant led to professional development workshops designed to encourage assignments that teach the effective use of numbers to frame an argument.

While we are confident in the usefulness of the rubric in this formative sense, we hope it will also prove useful in a summative context. The reliability results presented above suggest that raters at Carleton were able to achieve substantial reliability. In the future, we plan to test whether the rubric can be employed with similar reliability on other campuses including Wellesley College, Morehouse College, Iowa State University, and Edmonds Community College (Lynnwood, WA). The wide variety of institution types represented by this group will provide a good test of the broad applicability of the tool.

More research must also be done to establish construct validity. As Wallace et al. (2009, 11) quip, "a perfectly reliable ruler could be consistently wrong." We agree with those authors' assessment that the diversity of the QR concept means that we will not likely arrive at an "external gold standard"—an incontrovertible measure of QR against which assessment measures can be compared. But we can work to understand better how the conception of QR captured by the instrument presented here compares with that embedded in other assessment tools. For instance, James Madison University's Quantitative

Reasoning Test uses multiple-choice items to measure general education QR skill (Sundre 2008). Yet, the Council for Aid to Education (2008) asserts, "Life is not like a multiple choice test." Their Collegiate Learning Assessment (CLA) test asks students to respond in essay form to open-ended questions to a deeply contextualized case prompt. But for a few exceptions, most of the CLA prompts invite students to consider quantitative evidence. Examining the correlations between student scores on these alternative instruments and the QR-in-writing rubric might give us a better understanding of the various facets of QR and how they relate to one another.

Finally, the rubric presented here can help us understand better how students acquire QR facility. Do students with different majors achieve different levels of proficiency? Are some students more likely to compose QR-relevant arguments than others? How and when does QR use and proficiency develop over the undergraduate career? Do particular courses foster an appreciation for this important habit of mind? With a reliable assessment tool, we envision a robust research agenda answering these questions.

Acknowledgments

References

Bok, Derek. 2006. *Our underachieving colleges: A candid look at how much students learn and why they should be learning more.* Princeton, NJ: Princeton University Press.

Brakke, David F. 2003. Addressing societal and workforce needs. In *Quantitative literacy: Why numeracy matters for schools and colleges,* ed. Bernard L. Madison and Lynn Arthur Steen, 167−169. Princeton, NJ: National Council on Education and the Disciplines. http://www.maa.org/ql/pgs167_169.pdf (accessed Dec., 2009).

Camara, Wayne and Amy Schmidt. 2006. The new SAT: A comprehensive report on the first scores. Presented at College Board Forum 2006, November 10, 2006.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20(1): 37−46. http://dx.doi.org/10.1177/001316446002000104

College Board. 2003. How the new SAT essay will be scored.
http://info.gcsu.edu/intranet/univ_senate/Feb_23_04/Feb_04/newsat_scoring.pdf (accessed Dec. 7, 2009).

Council for Aid to Education. 2008. Collegiate learning assessment (CLA). CLA: Returning to learning. http://www.cae.org/content/pro_collegiate.htm (accessed Dec. 7, 2009).

De Lange, Jan. 2003. Mathematics for literacy. In *Quantitative literacy: Why numeracy matters for schools and colleges,* ed. Bernard L. Madison and Lynn Arthur Steen, 75−89. Princeton, NJ: National Council on Education and the Disciplines. http://www.maa.org/ql/pgs75_89.pdf (accessed Dec. 7, 2009).

Grawe, Nathan D. and B. Greg Marfleet. 2009. The use of quantitative reasoning across the curriculum: Empirical evidence from Carleton College. (Working paper). http://serc.carleton.edu/files/quirk/Assessment/grawe_marfleet_09.pdf (accessed Dec. 7, 2009).

Grawe, Nathan D. and Carol A. Rutz. 2009. Integration with writing programs: A strategy for quantitative reasoning program development. *Numeracy*, 2(2): Article 2. http://dx.doi.org/10.5038/1936-4660.2.2.2 (accessed Dec. 7, 2009).

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159−174. http://dx.doi.org/10.2307/2529310

Lutsky, Neil. 2008. Arguing with numbers: Teaching quantitative reasoning through argument and writing. In *Calculation vs. context: Quantitative literacy and its implications for teacher education,* ed. Bernard L. Madison and Lynn Arthur Steen, 59−74. Washington, DC: Mathematics Association of America. http://www.maa.org/ql/cvc/cvc-059-074.pdf (accessed Dec. 7, 2009).

Miller, Jane E. 2004. *The Chicago guide to writing about numbers*. Chicago: University of Chicago Press.

More or Less, British Broadcasting Corporation radio program. Retrieved April 27, 2007, from http://news.bbc.co.uk/2/hi/programmes/more_or_less/1628489.stm (accessed Dec. 7, 2009).

Richardson, Randall M. and William G. McCallum. 2003. The third R in literacy. In *Quantitative literacy: Why numeracy matters for schools and colleges,* ed. Bernard L. Madison and Lynn Arthur Steen, 99−106. Princeton, NJ: National Council on Education and the Disciplines. http://www.maa.org/ql/pgs99_106.pdf (accessed Dec. 7, 2009).

Schield, Milo. 2008. Quantitative literacy and school mathematics: Percentages and fractions. In *Calculation vs. context: Quantitative literacy and its*

implications for teacher education, ed. Bernard L. Madison and Lynn Arthur Steen, 87−107. Washington, DC: Mathematics Association of America. http://www.maa.org/ql/cvc/cvc-087-107.pdf (accessed Dec. 7, 2009).

Steen, Lynn Arthur, ed. 2001. *Mathematics and democracy: The case for quantitative literacy.* Washington, DC: Woodrow Wilson National Fellowship Foundation. http://www.maa.org/ql/mathanddemocracy.html (accessed May 29, 2009).

———. 2004. *Achieving quantitative literacy: An urgent challenge for higher education.* Washington, DC: Mathematical Association of America.

———. 2008. Reflections on Wingspread Workshop. In *Calculation vs. context: Quantitative literacy and its implications for teacher education,* ed. Bernard L. Madison and Lynn Arthur Steen, 11−23. Washington, DC: Mathematics Association of America. http://www.maa.org/ql/cvc/cvc-011-023.pdf (accessed Dec. 7, 2009).

Sundre, Donna. 2008. The Quantitative Reasoning Test, Version 9 (QR-9) Test Manual. The Center for Assessment & Research Studies. http://www.jmu.edu/assessment/resources/resource_files/QR-9 Manual 2008.pdf (accessed Dec. 7, 2009).

Taylor, Corrine. 2008. Preparing students for the business of the real (and highly quantitative) world. In *Calculation vs. context: Quantitative literacy and its implications for teacher education,* ed. Bernard L. Madison and Lynn Arthur Steen, 109−124. Washington, DC: Mathematics Association of America. http://www.maa.org/ql/cvc/cvc-109-124.pdf (accessed Dec. 7, 2009).

———. 2009. Assessing quantitative reasoning. *Numeracy*, 2(2): Article 1. http://dx.doi.org/10.5038/1936-4660.2.2.1 (accessed Dec. 7, 2009).

Wallace, Dorothy, Kim Rheinlander, Steven Woloshin, and Lisa 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 Dec. 7, 2009).

Wiggins, Grant. 2003. 'Get real!': Assessing for quantitative literacy. In *Quantitative literacy: Why numeracy matters for schools and colleges,* ed. Bernard L. Madison and Lynn Arthur Steen, 121−143. Princeton, NJ: National Council on Education and the Disciplines. http://www.maa.org/ql/pgs121_143.pdf (accessed Dec. 7, 2009).

Appendix: Suggestions for Creating Similar Rubrics

The rubric presented here has been developed and revised over four years. The reliability of early versions was tested by a single pair of readers. These readers achieved roughly 80% agreement in a reading of around 100 papers. Following some further revision, the rubric was tested by a group of about a dozen readers. The larger group came to similarly strong levels of agreement when assessing relevance and extent of QR. But evaluations of the quality of implementation, interpretation, and communication (three separate scores in that version of the rubric) were far less reliable. Another round of revision led to the current form of the rubric.

Recognizing that others seeking to assess QR in argument may have somewhat different objectives or student populations, we expect that adaptation may require rubric revision. Below we note several lessons we learned during rubric development that may facilitate this adaptation elsewhere.

Less is more. As mentioned above, the original rubric asked raters to assess three distinct elements of QR quality: implementation, interpretation, and communication. Discussions during norming sessions suggested that readers had a difficult time distinguishing between these intertwined concepts. Our current practice of requiring a single holistic score eliminated these challenges.

Similarly, the original rubric provided a greater range of scores for both extent and quality of QR. As our discussions progressed we realized that some disagreements arose simply because the number of scores exceeded the number of categories readers had in mind. A reduction in scoring levels eliminated more or less arbitrary scoring decisions.

More is more. While we simplified the scoring range, we substantially expanded the codebook language used to describe scoring distinctions. An explicit scoring matrix put in writing the discussions held during norming sessions.

Norming matters. No matter how clear the codebook and scoring sheet, effective norming sessions remain critical. While the scoring matrix ensures we are all using the same language to describe our ratings, discussions during norming sessions revealed important differences in raters' interpretation of that language. We have found that about two hours are needed for the discussion of the codebook and a list of common paper types and to read and discuss a common set of (carefully chosen) papers. This investment easily repays its cost.

Order issues. Readers had strong preference as to the ordering of items on the scoring sheet. Because we are trying to read papers from a "neutral" perspective without regard for the nature of the assignment or the department for which the paper was written, raters preferred that they not be asked to consider the assignment until after scoring the paper. In fact, one reader asked if the pages

might be arranged in the future such that the cover sheet (which includes the course number) and assignment follow the paper. This seems like a good suggestion.

Similarly, in an earlier version of the rubric the coding of problem characteristics preceded holistic quality assessment. Several raters found this distracting. They pointed to papers which seemed to be good (though not exemplary) in a holistic sense that nevertheless exhibited several problematic characteristics in one place or another. Having just coded for the presence of problem characteristics, these readers found it hard to give the paper the sound holistic score they felt it deserved. While the rubric was revised to meet this request, it seems this change may have reduced reliability in the assessment of problem characteristics. It may be easier to code for these issues as they occur rather than to try and recall them after reading the entire paper. (On the other hand, if the intention is only to flag problems that "significantly detract from the argument," a reader's recall difficulty may be a good thing.) Whatever the merit of this change, it is clear that the order of rubric items matters and should be considered carefully.