**Analysis of Feasibility Survey**

Table 1 presents a summary of the feasibility survey completed by the 7 scorers who took part in the exercise. All questions were rated on a scale of 1 (strongly disagree) to 6 (strongly agree). In questions 1-4 respondents were asked to evaluate the connection between QR and argument in general—not specifically with our rubric. In questions 5-13 they were asked to respond to the particulars of our rubric.

 Responses to question 1 show nearly unanimous strong agreement that communication is a key element of QR. Respondents reported similarly strong agreement that successful QR education should be evidenced in student written work (question 3) and thus that we should look to see how QR is or is not playing out in student papers (question 4). However, while generally agreeing, respondents were less confident that our rhetorical approach would draw in colleagues from ALH (question 2). This may in part reflect the fact that we had few ALH folks in the group and so it would be hard to assess whether the approach was appealing to that population.

**Table 1. Summary of agreement/disagreement responses to feasibility survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean | Mode | 75th Percentile | Median | 25th Percentile |
| *1. Communication plays a significant role in what it means to be quantitatively literate.* |
| 5.857143 | 6 | 6 | 6 | 6 |
| *2. Emphasizing the rhetorical aspects or approach to QR increases the likelihood that colleagues in the arts, literature, and humanities will teach QR practice.*  |
| 4.428571 | 4 | 5 | 4 | 4 |
| *3. If we are succeeding in our efforts to enhance students’ quantitative reasoning skills, we should see evidence of that in student written work.* |
| 5.571429 | 6 | 6 | 6 | 5 |
| *4. For a complete understanding of how well our students are doing in the area of QR, we should consider looking for evidence in student writing in addition to traditional quantitative assessment tests.* |
| 5.714286 | 6 | 6 | 6 | 5.5 |
| *5. Applying the rubric gave me important insights into how students are (or are not) using quantitative evidence in written arguments.* |
| 4.571429 | 5 | 5 | 5 | 4.5 |
| *6. After reading student work with the rubric, I have an idea for how I might alter one or more assignments to help students improve their QR skills.* |
| 4.857143 | 6 | 6 | 5 | 4 |
| *7. Overall, I found the rubric easy enough to employ.* |
| 3.428571 | 3 | 4.5 | 3 | 2.5 |
| *8. I was able to distinguish between the categories of QR relevance.* |
| 4.714286 | 4 | 5.5 | 5 | 4 |
| *9. I was able to distinguish between the categories of QR extent.* |
| 5.142857 | 6 | 6 | 5 | 4.5 |
| *10. I was able to distinguish between the levels of quality on papers which I deemed peripherally relevant.* |
| 4.666667 | 4 | 5 | 4.5 | 4 |
| *11. I was able to distinguish between the levels of quality on papers which I deemed to be centrally relevant.* |
| 4.285714 | 4 | 5 | 4 | 4 |
| *13. It would be possible to use a rubric like this one to assess student work at Wellesley.* |
| 5.142857 | 5 | 5.5 | 5 | 5 |
| *14. I would be useful to use a rubric like this one to assess student work at Wellesley.* |
| 4.857143 | 6 | 6 | 5 | 4 |

 Turning next to the questions which related specifically to our rubric, respondents generally agreed that its application gave them important insights into student work (question 5) and that these insights were actionable—that they suggested ways in which assignments might be revised to improve QR on campus (question 6). This confirms Carleton’s experience with the rubric as a successful formative tool.

 While scorers did not generally find the rubric easy to use (question 7), they reported that they were able to do so (questions 8-11). Interestingly, the most difficultly appears to have come in assessing the quality of centrally relevant papers. Qualitative responses suggest the difficulty was related to the fact that the rubric includes only 4 quality categories. Scorers wanted more. (This may also be an issue at Carleton where science papers seem to receive a disproportionate share of 4s.) While it isn’t certain why scorers did not use more 3s to provide more scoring variation, one possible explanation is that they used 3s for central papers outside the sciences and then lacked a means of differentiating among sciences papers, all of which were deemed better than the non-science central QR papers. If that is the case, then we may have an issue to resolve as it would seem that genre should not be so key in determining the quality of QR in argument.

 Ultimately, scorers strongly agreed that it would be feasible to employ the rubric at Wellesley (question 12). With a means score over 5 and no response lower than 5, participants reached consensus on this point. When asked if it would be useful to do so (question 13), agreement was nearly as strong (mean just under 5 and only 1 person selecting a score less than 4).

**Analysis of Assessment Data**

*Limited Reliability Analysis*

Only 23 papers were read by two readers so it is difficult to precisely estimate reliability—especially when looking at a subset of papers like “centrally relevant” or even when excluding one of the seven scorers. For what it is worth in this small sample, readers reached exact agreement about relevance 87.0% of the time (Cohen’s κ = 0.77). Slightly lower agreement was found in extent of QR (76.2% agreement or κ = 0.50). This all looks good.

Unfortunately, readers did not agree well over quality (16.7% agreement or κ = -0.21). Table 2 shows the joint distribution of quality scores. Not only do the readers achieve less exact agreement than one would expect by chance alone, they actually show a strong negative correction (-0.64) which is statistically significant! It’s hard to make too much of this with the small sample of readers. When scorers are removed one at a time, agreement sometimes increased. But with such a small sample size it is hard to make much an argument for excluding any one. It also may be that the science-centric sample made it difficult for some readers to assess quality given their personal stats background. I’m not sure if I mentioned in norming that it was okay to put a paper back. That may have been a mistake (but then again, given the sample, there may not have been much we could do about it.) In total, the low quality reliability seems to be raise questions. Is it not possible to achieve a common perspective in such a short time? Is it not possible to have a cross-cutting group read largely science papers?

**Table 2. Distribution of quality scores given to papers read by two readers.**

|  |  |
| --- | --- |
| **Score Given by**  | **Score Given by Reader 2** |
| **Reader 1** | **1** | **2** | **3** | **4** |
| **1** | 0 | 0 | 1 | 1 |
| **2** | 0 | 0 | 3 | 2 |
| **3** | 3 | 3 | 3 | 0 |
| **4** | 0 | 2 | 0 | 0 |

*Assessment Analysis*

Table 3 presents data on how frequently student papers were deemed to be QR relevant. The science-heavy sample is clearly evident with fewer than 10% of the papers found to be non-QR relevant. Clearly, the assignment mattered. When explicit QR was called for, students always took an approach for which QR was relevant and nearly 98% of the time it mattered centrally. Even in papers that did not explicitly call for quantitative evidence, almost all students wrote papers for which QR was “in play.” This may reflect students perceptions of faculty members expectations in science courses.

**Table 3. Relevance of QR by assignment type**

|  |  |
| --- | --- |
|  | Relevance |
|  | None | Peripheral | Central |
| All Papers | 6.7% | 10.0% | 73.3% |
|  |  |  |  |
| Assignment did not require QR | 13.7% | 39.2% | 47.1% |
| Assignment required QR | 0.0% | 2.7% | 97.6% |
|  |  |  |  |

 Table 4 considers the extent of QR found in the papers. Overall, this was a very QR-y group of papers. Nearly three-fourths employed quantitative evidence throughout the paper. As expected, when QR was not relevant to a paper, students did not employ QR. What may be notable is the propensity of Wellesley students to use QR when relevant. When centrally relevant, 90% of students use QR extensively and another 8% do some. By contrast, in Carleton’s sample roughly 20% of students use no QR when it is centrally relevant to the paper. Similarly, a relatively modest 18.5% of students failed to use any QR when peripherally relevant. At Carleton the comparable figure is 95%. Here again, the nature of the sample might matter. A similar pattern is evident in the data by assignment type: even when not required to, Wellesley students called on quantitative evidence with fair regularity. Students may be more prone to using QR in science courses than in general education courses or they may learn to do so with maturity. While we cannot be certain as to the causes for these differences, it seems it would be valuable for us to find them out.

**Table 4. Extent of QR by paper and assignment type**

|  |  |
| --- | --- |
|  | Extent |
|  | None | Some | Extensive |
| All Papers | 11.8% | 14.7% | 73.5% |
| *Paper type* |  |  |  |
| QR-Irrelevant | 100% | 0.0% | 0.0% |
| Peripherally Relevant | 18.5% | 44.4% | 37.0% |
| Centrally Relevant | 2.0% | 8.0% | 90.0% |
| *Assignment type* |  |  |  |
| Assignment did not require QR | 23.5% | 27.4% | 49.0% |
| Assignment required QR | 1.4% | 4.2% | 94.4% |
|  |  |  |  |