Closing the Loop: Involving Faculty in the Assessment of Scientific and Quantitative Reasoning Skills of Biology Majors

By Carol A. Hurney, Justin Brown, Heather Peckham Griscom, Erika Kancler, Clifton J. Wigtil, and Donna Sundre

The development of scientific and quantitative reasoning skills in undergraduates majoring in science, technology, engineering, and mathematics (STEM) is an objective of many courses and curricula. The Biology Department at James Madison University (JMU) assesses these essential skills in graduating biology majors by using a multiple-choice exam called the Natural World-9 (NW-9). NW-9, comprised of measures of Quantitative and Scientific Reasoning, contains items developed by faculty at JMU to assess the impact of the General Education program on the development of scientific and quantitative reasoning skills in a content-independent manner. We discuss methodology we used to involve faculty in determining the generalizability of NW-9 to assess the objectives of the biology curriculum and setting standards to interpret student achievement on NW-9. Student performance on NW-9 identified both strong and weak areas in our instruction and suggested that our biology faculty needs to reevaluate methodology for teaching students how to interpret and analyze data. More important, we can close the assessment loop by allowing faculty to participate in the assessment process and meaningfully reflect on student assessment results.

There are three options available to faculty interested in assessing the impact of undergraduate education on scientific and quantitative reasoning skills: use an existing instrument, modify an existing instrument, or develop a new instrument. Given the importance that science, technology, engineering, and mathematics (STEM) programs and national science organizations place on the development of scientific and quantitative reasoning skills, one would expect to find an endless array of reliable instruments that assess whether students graduating from undergraduate programs successfully acquired these essential skills (Howard Hughes Medical Institute 1996; NRC 2003). Many of the standardized tests, such as the Graduate Record Examination, include items that assess scientific reasoning ability, but for the most part research-based standardized tests address content knowledge (Bao et al. 2009). The Classroom Test of Scientific Reasoning developed by Lawson in 1978 is still popular among STEM educators, but this instrument addresses very broad areas of scientific reasoning and does not assess quantitative reasoning skills (Lawson 1978). Unfortunately, few readily accessible instruments are available that reliably assess both scientific and quantitative reasoning skills in undergraduates.

James Madison University (JMU) is a publicly funded, comprehensive institution of approximately 18,000 students in Harrisonburg, Virginia and has a strong emphasis on program assessment. The nationally recognized Center for Assessment and Research Studies (CARS) provides significant resources to the development of a nationally recognized assessment program (*www.jmu.edu/assessment/*). Building on the need for assessment of scientific and quantitative reasoning in higher education, and more specifically to inform STEM education, members of CARS in partnership with JMU faculty developed the Natural World-9 (NW-9) instrument, which contains two components: the Scientific Reasoning Test (SR-9; Sundre, 2008) and the Quantitative Reasoning Test (QR-9; Sundre, Thelk, and Wigtil 2008). All NW-9 items were written by James Madison University science and mathematics faculty to assess the objectives of the science component of the General Education program (see Table 1). Rather than investing faculty time in developing a new instrument, we decided to explore whether the NW-9 instrument developed and tested by CARS could assess scientific and quantitative reasoning skills in biology majors. We also wanted to involve faculty in this process to enhance faculty understanding and appreciation of the assessment process and results.

The Department of Biology has 56 full-time and part-time faculty, approximately 900 declared majors, and 100–125 students who graduate

each year. The biology curriculum is designed upon an explicit set of content, skill, and experience learning objectives developed by biology faculty. These objectives support the two major goals of the curriculum: insuring that biology majors are literate in the scientific process and integrating research experiences into the learning environment for all our majors. Specifically, the skill objectives concentrate on scientific reasoning skills (see Table 1, skill objectives 1–10), but they also include objectives related to effective communication skills (see Table 1, skill objectives 11–14) and the ability to use quantitative reasoning skills to analyze biological phenomena (see Table 1, skill objectives 7 and 14). Assessment of the skill objectives is based on the results of two instruments, a modified version of the Academic Skills Inventory (ASI; Kruger and Zechmeister 2001) and the NW-9. The ASI differs from the NW-9 instrument in that the ASI asks students to report their experience level with a variety of academic skills, whereas the NW-9 instrument directly measures skill level. Results from the ASI indicate that students self-report behavioral gains in skills associated with written and oral communication, research methodology, and statistics (Seifert et al. 2009). Although the ASI provides insights regarding how well graduates of the biology major achieve some of the skill objectives, the NW-9 exam provides a more direct measurement of scientific and quantitative reasoning skills.

Although the NW-9 instrument was designed to assess the General Education learning objectives, there are many features of NW-9 that suggest this instrument will provide meaningful data to assess the skill objectives of the biology major. First, many of the General Education objectives are similar to the biology major skill objectives. For example, skill objectives 7, 9, and 10 and General Education objective 8 both discuss the ability of students to evaluate scientific sources, and skill objective 1 and General Education objective 6 both explore students' ability to distinguish between association and causation. Second, CARS has extensively tested both components of NW-9 to establish two important measures of a meaningful assessment instrument: reliability and validity. The NW-9 instrument reliability and validity scores suggest that the instrument consistently measures the scientific and quantitative reasoning objectives of the General Education program (Sundre 2008; Sundre, Thelk, and Wigtil 2008). Third, NW-9 items do not test specific content knowledge. Rather, many of the items provide content necessary to determine the answer (see Figure 1a), whereas other items test concepts that do not rely on factual information (see Figure 1b). Based on these features of NW-9, we determined the generalizability of the NW-9 instrument to assess the skill objectives of the biology major. We did this by involving biology faculty in a content alignment process in which they mapped NW-9 items to the skill objectives. We also involved faculty in the standard setting protocol to determine the standards for acceptable performance of our graduating biology seniors on items that mapped to the skill objectives. Results from these endeavors allow us to (1) evaluate senior biology major students' performance on the mapped items; (2) determine whether students fell below, met, or exceeded faculty standards; and (3) discuss NW-9 assessment results at

TABLE 1

Comparison of biology major skill objectives (*N* = 14) with General Education Cluster 3 objectives (*N* = 7).

Biology major skill objectives

1. Discriminate between association and causation, and identify the types of evidence used to establish causation.
2. Formulate a hypothesis and identify relevant variables necessary to test that hypothesis.
3. Design and execute experiments to test hypotheses.
4. Obtain data.
5. Organize data.
6. Analyze and interpret data.
7. Evaluate a statement, hypothesis, or claim using numerical or other evidence.
8. Locate sources of scientific information.
9. Evaluate the reliability of sources.
10. Critically evaluate a paper from the primary scientific literature.
11. Use effective professional communication in posters.
12. Use effective professional communication in lab reports.
13. Use effective professional communication in oral reports.
14. Use mathematics to understand and analyze biological phenomena.

General Education Cluster 3 objectives

1. Describe the methods of inquiry that lead to mathematical truth and scientific knowledge and be able to distinguish science from pseudoscience.
2. Use theories and models as unifying principles that help us understand natural phenomena and make predictions.
3. Recognize the interdependence of applied research, basic research, and technology, and how they affect society.
4. Illustrate the interdependence between developments in science and social and ethical issues.
5. Use graphical, symbolic, and numerical methods to analyze, organize, and interpret natural phenomena.
6. Discriminate between association and causation, and identify the types of evidence used to establish causation.
7. Formulate hypotheses, identify relevant variables, and design experiments to test hypotheses.
8. Evaluate the credibility, use, and misuse of scientific and mathematical information in scientific developments and public-policy issues.

departmental retreats regarding pedagogical strategies utilized by biology faculty to address the skill objectives.

Methods

Content alignment of NW-9 items to skill objectives

A critical step in determining the generalizability of the NW-9 instrument is to examine the content alignment between test items and skill objectives (D'Agostino et al. 2008). The degree of content alignment determines the ability of individual items to provide accurate information on student performance for each objective. Based on advice from the assessment experts at the CARS, we utilized item-level analysis to determine content alignment of the NW-9 instrument to the skill objectives (Martone and Sireci 2009). This was accomplished by recruiting eight faculty members, representing various subdisciplines in biology, to analyze the 66 items on the NW-9 instrument. Each faculty member provided independent judgments on whether an item successfully assessed one or more of the skill objectives. Faculty members were asked to review one stated learning objective at a time and determine whether or not each NW-9 item successfully assessed that objective. A dichotomous choice was provided for each item (*yes* or *no*). After making judgments about one objective, the faculty member proceeded to the next skill objective. No additional discussions or attempts to form consensus were attempted. This objective by objective procedure is less arduous for faculty than attempting to simultaneously make judgments about individual items across all learning objectives (D'Agostino et al. 2008). In consultation with the CARS, we developed a fairly stringent rule that an item would be deemed successfully mapped to a skill objective if six out of the eight evaluators (75%) assigned the item to a particular objective.

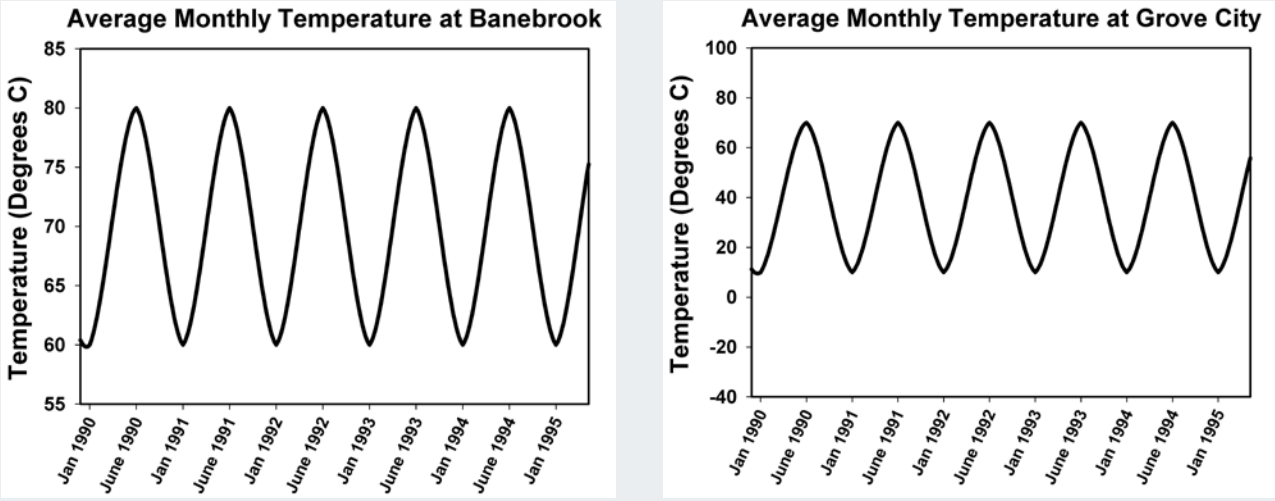
Establishing faculty standards

We used a modified Angoff method to establish a faculty standard for

FIGURE 1

Two examples of NW-9 items: (a) question that requires students to demonstrate proficiency in more than one skill, and (b) question that assesses the ability of students to interpret data.

(a) Regarding the two graphical displays given below, which of the following statements is correct?



a. Banebrook has the largest changes in temperature throughout the year.
b. Banebrook and Grove City temperatures exhibit exponential behavior throughout the year.
c. Neither of the above.

(b) Suppose a researcher wants to test the hypothesis that exposure to cadmium in childhood causes neurological damage that reduces IQ. The researcher randomly selects 500 fourth graders, monitors their cadmium exposure for one year, and then tests each student's IQ. The researcher finds that as cadmium exposure increases, IQ declines. Can the researcher conclude from the observed association between cadmium exposure and intelligence that cadmium causes reduced IQ?

a. No. The researcher did not include enough persons in the study.
b. No. There may be a third variable associated with exposure to cadmium that actually causes the lowered IQ.
c. Yes. The researcher followed the scientific method.
d. Yes. An association between the amount of cadmium exposure and lowered IQ is exactly what we would predict from the hypothesis.

each skill objective to provide greater interpretive power regarding student results (Maurer et al. 1991). The Angoff method provides a quantitative benchmark to determine whether graduating seniors are meeting faculty expectations. Biology faculty members ($n = 15$) who had no knowledge of student test performance examined each of the NW-9 items that mapped to the skill objectives. The faculty volunteers were asked to provide a judgment of the percentage of graduating biology majors who should provide a correct response for each item. During this exercise, faculty members were asked to not discuss their ratings until after completion of the entire exercise. Following Angoff methodologies, faculty ratings for each item were grouped, on the basis of the mapping data, to the appropriate skill objectives. The mean of the scores for each skill objective represents the faculty standard for student success (see Table 2).

Determining student performance on NW-9

We administered the NW-9 instrument to 214 graduating seniors (88 in 2008 and 126 in 2009). The mean student scores on the suite of questions corresponding to each of the seven skill objectives were calculated and transformed to the percentage correct. For each objective, the faculty standards were compared with the performance of the graduating seniors using a Mann-Whitney U nonparametric test with sequential Bonferroni post hoc analysis (see Table 2). Cohen's *d* was used to determine effect size. If the mean student score for an objective was significantly higher than the faculty standard, students exceeded the faculty standard for that objective. If the mean student scores were not significantly different from the faculty standard, then students met the faculty standard. If the mean student score was significantly lower than the faculty standard, then students did not meet the faculty standards.

Results

Content alignment of NW-9 items to the skill objectives

The stringent content alignment activity we utilized revealed that 25 of the 66 items strongly mapped to 7 of the 14 skill objectives. The objectives for which items were successfully aligned relate to distinguishing association from causation, formulating and evaluating hypotheses, designing experiments, analyzing and interpreting data, and using mathematics to understand biological phenomena (see Table 2). We found that multiple items were assigned to each of these seven objectives. However, using the established criteria, there were no items that mapped to skill objectives relating to obtaining data; organizing data; locating sources of scientific information; evaluating the reliability of sources; critically evaluating a paper from the

TABLE 2

Number of NW-9 items mapped, faculty standard, and student performance for six skill objectives.

Skill objective	NW-9 items mapping to objective	Faculty standard	Student performance
Student performance exceeded faculty standard			
3. Design and execute experiments to test hypotheses.	3 items (5% of test)	84.8%	91.6% ($p < .0001, d = .27$)
14. Use mathematics to understand and analyze biological phenomena.	2 items (3% of test)	74.8%	87.1% ($p < .0001, d = .36$)
Student performance met faculty standard			
1. Discriminate between association and causation, and identify the types of evidence used to establish causation.	6 items (9% of test)	79.3%	75.5% ($p = .5920, d = .53$)
2. Formulate a hypothesis and identify relevant variables necessary to test that hypothesis.	11 items (17% of test)	82.6%	86.5% ($p = .050, d = .21$)
7. Evaluate a statement, hypothesis, or claim using numerical or other evidence.	15 items (23% of test)	78.3%	75.7% ($p = .9740, d = .17$)
Student performance fell below faculty standard			
6. Analyze and interpret data.	23 items (33% of test)	81.0%	70.4% ($p < .009, d = .58$)

Note: The faculty standard was derived from biology faculty predicting the percentage of graduating biology majors whom they thought would provide a correct response for each item (refer to *Faculty standards* in the Results section). Student performance was the percentage of correct answers that mapped to each skill objective. For each objective, the faculty standards were compared with the performance of the graduating seniors using a Mann-Whitney U nonparametric test with sequential Bonferroni post hoc analysis. Cohen's *d* was used to determine the effect size (*d*). NW-9 = Natural World-9.

primary literature; and using effective professional communication in posters, lab reports, and oral reports (skill objectives 4, 5, 8–13). This was an expected and validated finding. These learning objectives are not amenable to selected response item types. We currently use the ASI and are exploring other more direct methods to assess these skills and competencies. Some of the other NW-9 items that did not map to the skill objectives are designed to assess General Education objectives that do not align with faculty-developed curricular objectives of the biology major, such as understanding the difference between basic and applied research.

The most highly assessed objective was skill objective 6, analyzing and interpreting data, as 33% of the NW-9 items mapped to this objective (see Table 2). An example of a NW-9 item that assesses the ability of students to interpret data is shown in Figure 1a. Content in this item is not directly addressed in any biology course, which allows us to determine whether the student can transfer and generalize knowledge to interpret data in a situation in which they are not familiar with the content. Many items mapped to more than one skill objective, which reflects that many of the NW-9 items require students to demonstrate proficiency in more than one skill to achieve the correct answer. Overall, the content alignment activity provided validation for the use of NW-9 test scores to assess many of the quantitative and scientific reasoning objectives of our curriculum.

Faculty standards

For the most part, the faculty standards for each skill objective were in the 75%–85% range (see Table 2). The highest faculty standard, 84.8%, was for designing and executing experiments to test hypotheses, whereas the lowest, 74.8%, was for using mathematics to understand biological phenomena.

Student performance

Graduating biology majors exceeded faculty expectations for two skill objectives, met faculty expectations for three skill objectives, and fell below faculty standards for one skill objective (see Table 2). Seniors exceeded the faculty standard for designing experiments and using mathematics to understand a biological phenomena (skill objectives 3 and 14). In particular, the average score for items that map to designing and executing experiments (skill objective 3) was 91.6%, which is much higher than the faculty standard of 84.8% ($p = .0001$, $d = .36$). Graduating seniors met the faculty standard for formulating hypotheses, discriminating between association and causation, and evaluating a statement or claim using evidence (skill objectives 1, 2, and 7). Finally, our assessment results indicate that seniors correctly answered 70.4% of the questions that map to skill objective 6, which is significantly lower than the faculty standard for analyzing and interpreting data (81.0%, $p < .009$, $d = .58$).

Discussion

As a result of this project, we have empirical evidence that the NW-9 provides meaningful measures of quantitative and scientific reasoning skills in biology majors. We found that 25 items on the NW-9 instrument map to seven of the skill objectives of the biology curriculum. The curriculum objectives assessed by NW-9 represent essential scientific and quantitative reasoning skills. Most notably, the exam scores provide insight into students' abilities to identify and evaluate evidence that can be used to establish causation, formulate hypotheses, identify relevant variables to test hypotheses, analyze and interpret data, and use mathematics to understand biological phenomena. We recognize that the skill objectives not assessed by NW-9 are difficult to evaluate with a multiple-choice exam (e.g., effectiveness in presenting scientific research), and we will seek new direct methodologies.

Faculty had an overall prediction that on average, 79% of the seniors would answer correctly the suite of NW-9 questions that mapped to the skill objectives. Faculty expectations across the objectives showed some variability, ranging from approximately 75%–85% correct. Some items and objectives were determined to be more challenging than others. Actual student performances ranged from approximately 70%–92% correct for a suite of questions mapped to a particular objective. Faculty expectations were highest (84.8% expected to answer correctly) for questions that mapped to the skill objective related to designing and executing experiments to test hypotheses. Student performance was also highest for questions that mapped to this objective (91.6% of the students answered these questions correctly).

Faculty standards were relatively high (>82.6% of students were predicted to answer the question correctly) for NW-9 questions referring to the skill objective of formulating hypotheses and designing and executing experiments to test hypotheses (see Table 2). This may be because experimental design is emphasized in biology courses, thus faculty members have higher expectations for these skills. The faculty standard was lowest for the objective related to using mathematics to answer biological phenomena (<75% of students were predicted to answer the questions correctly; see Table 2). This may be related to the difficulty of items that assess quantitative skills, but it could also reflect that faculty members do not feel confident that courses in the biology curriculum address these skills. Likewise, faculty standards for students' ability to analyze and interpret data were on the low end of the spectrum (81%). This is surprising given that many laboratory courses emphasize the use of statistics to analyze data.

We found that the NW-9 exam can be used to assess many of the JMU Biology Department skill objectives, which are most likely similar to the objectives other Biology Departments have for their students. Our results demonstrate that the NW-9 exam can be used to assess scientific and quantitative reasoning skills in areas outside of the General Education curriculum. Institutions interested in implementing instruments, such as NW-9, should map the items to their curriculum objectives and set faculty standards, as these will vary with student populations, curriculum, and faculty expectations. Once student performance data is collected, faculty can identify areas of strength and weakness in instruction and/or curriculum.

Overall, results from the NW-9 instrument in conjunction with the results from the ASI (Seifert 2009) suggest that the current biology major curriculum produces students who have met or exceeded faculty expectations for most of the specified curriculum skill objectives. We also noted a weak area in the curriculum regarding the skill of analyzing and interpreting data. This suggests a need for conversations to occur between laboratory instructors in regards to this essential skill objective. Laboratory courses should be targeted, because this is where the majority of inquiry-based learning occurs, such as analyzing and interpreting data. This study provides a baseline measure for the impact of the curriculum on skill development. We will continue to monitor our assessment results to measure the impact of changes we implement in laboratory courses to see if these changes increase student skill in data analysis.

One of the most significant outcomes we observed as we implemented our assessment design was an increase in faculty participation and interest in the assessment process and student results. By involving biology faculty in the content alignment and standard setting activities, we created a customized process that we can use, as a department, to analyze student performance in the areas of scientific and quantitative reasoning. More important, we have created a culture of assessment in our department that reflects the goals of the curriculum, the perspective of the faculty, and an awareness of student learning outcomes. This process has helped us to "close the loop" with understanding and using our assessment results. Our faculty conversations about assessment, our program, and our students' learning have been deepened and enriched. Most important, these results provide our faculty with compelling evidence that the NW-9 instrument measures many of the biology-major student learning objectives. We were able to engage many of our faculty in the development of a community-established expectation for student performance. Finally, this set of student performance expectations gave us a new and valued interpretive framework for our assessment results. ■

References

Bao, L., T. Cai, K. Koenig, K. Fang, J. Han, J. Wang, Q. Liu, et al. 2009. Learning and scientific reasoning. *Science* 323 (5914): 586–587.

D'Agostino, J.V., M.E. Welsh, A.D. Cimetta, L.D. Falco, S. Smith, W.H. VanWinckle, and S.J. Powers. 2008. The rating and matching item-objective alignment methods. *Applied Measurement in Education* 21 (1): 1–21.

Howard Hughes Medical Institute. 1996. *Beyond Bio 101: The transformation of undergraduate biology education*. Chevy Chase, MD: Howard Hughes Medical Institute. *www.hhmi.org/BeyondBio101*

Kruger, D.J., and E.B. Zechmeister. 2001. A skills-experience inventory for the undergraduate psychology major. *Teaching of Psychology* 28 (4): 249–253.

Lawson, A.E. 1978. The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching* 15 (1): 11–14.

Martone, A., and S.G. Sireci. 2009. Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research* 79 (4): 1332–1361.

Maurer, T.J., R.A. Alexander, C.M. Callahan, J.J. Bailey, and F.H. Dambrot. 1991. Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personal Psychology* 44 (2): 235–262.

National Research Council (NRC). 2003. *Biology 2010: Transforming undergraduate education for future research biologists*. Washington, DC: National Academies Press.

Seifert, K., C.A. Hurney, C.J. Wigtil, and D.L. Sundre. 2009. Using the academic skills inventory (ASI) to assess the biology major. *Assessment Update* 21 (1–2): 14–15.

Sundre, D.L. 2008. *The Scientific Reasoning Test, Version 9 (SR-9) test manual.* Harrisonburg, VA: Center for Assessment and Research Studies. *www.madisonassessment.com/assessment-testing/scientific-reasoning-test/*

Sundre, D.L., A. Thelk, and C. Wigtil. 2008. *The Quantitative Reasoning Test, Version 9 (QR-9) test manual.* Harrisonburg, VA: Center for Assessment and Research Studies. *www.madisonassessment.com/assessment-testing/quantitative-reasoning-test/*

Carol A. Hurney *is an associate professor of biology and executive director of the Center for Faculty Innovation,* ***Justin Brown*** *is an assistant professor of biology,* ***Heather Peckham Griscom*** (griscohp@jmu.edu) *is an associate professor of biology, and* ***Erika Kancler*** *is an assistant professor of biology, all at James Madison University (JMU) in Harrisonburg, Virginia.* ***Clifton J. Wigtil*** *is graduate student in gifted education at Purdue University.* ***Donna Sundre*** *is a professor of psychology and the executive director of the Center for Assessment and Research Studies at JMU.*