4-23-2010

Development of a 'Universal' Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing

Briana Eileen Timmerman
University of South Carolina - Columbia, timmerman@sc.edu

Denise Strickland
University of South Carolina - Columbia

Robert L. Johnson
University of South Carolina - Columbia, rjohnson@mailbox.sc.edu

John R. Payne
University of South Carolina - Columbia

Publication Info

Postprint version. Published in *Assessment & Evaluation in Higher Education*, Volume 36, Issue 5, 2010, pages 509-547.

Assessment & Evaluation in Higher Education
2010, 1–39, iFirst Article

Routledge
Taylor & Francis Group

Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing

Briana E. Crotwell Timmerman[a,b]*, Denise C. Strickland[c], Robert L. Johnson[c] and John R. Payne[c]

[a]SC Honors College, University of South Carolina, Columbia, SC 29208, USA; [b]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA; [c]Department of Educational Studies, University of South Carolina, Columbia, SC 29208, USA

We developed a rubric for measuring students' ability to reason and write scientifically. The Rubric for Science Writing (Rubric) was tested in a variety of undergraduate biology laboratory courses (total n = 142 laboratory reports) using science graduate students (teaching assistants) as raters. Generalisability analysis indicates that the Rubric provides a reliable measure of students' abilities (g = 0.85) in these conditions. Comparison of student performance in various biology classes indicated that some scientific skills are more challenging for students to develop than others and identified a number of previously unappreciated gaps in the curriculum. Our findings suggest that use of the Rubric provides three major benefits in higher education: (1) to increase substance and consistency of grading within a course, particularly those staffed by multiple instructors or graduate teaching assistants; (2) to assess student achievement of scientific reasoning and writing skills; and (3) when used in multiple courses, to highlight gaps in alignment among course assignments and provide a common metric for assessing to what extent the curriculum is achieving programmatic goals. Lastly, biology graduate students reported that use of the Rubric facilitated their teaching and recommended that training on the Rubric be provided to all teaching assistants.

Keywords: rubric; science writing; laboratory reports; programmatic assessment; scientific reasoning

Motivation and rationale

A fundamental goal of science education is that students develop the ability to actually *do* science (American Association for the Advancement of Science 1993; National Research Council 1996). For coursework and the grades earned therein to reflect whether or not students are gaining these skills, faculty and institutions of higher education need meaningful measures of student ability and performance in authentic scientific tasks (Cabrera, Colbeck, and Terenzini 2001; Shavelson and Huang 2003). Attempting to measure scientific inquiry skills in a meaningful fashion using realistic tasks such as oral presentations of scientific work is not new. Published performance measures exist for online discussions (Clark and Sampson 2006), lab notebooks (Baxter et al. 1992; Ruiz-Primo et al. 2004), verbal discussions (Erduran, Simon, and Osborne 2004; Hogan, Nastasi, and Pressley 2000), oral presentations (Hafner and Hafner 2003), and direct observation and scoring of students as they work in the laboratory

*Corresponding author. Email: timmerman@schc.sc.edu

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930903540991
http://www.informaworld.com

(Baxter et al. 1992; Germann and Aram 1996). Many of these venues are not traditionally assessed in large university classes, however, due to the limitations of instructor time and resources (Alters and Nelson 2002). To assess gains in student learning over multiple courses, we also needed to select a venue that was common to a variety of courses. Written exams and laboratory reports were the most common means of assessment in our courses.

We selected the laboratory report as a data source because it seemed more likely than written exams to be a rich source of information about students' scientific reasoning skills. Additionally, the merit of professional scientists is often assessed by their success in receiving grant funding and/or publishing scientific findings. As both of these involve translating scientific reasoning and inquiry skills into scientific writing, we selected laboratory reports as a reasonable proxy for measuring similar scientific skills in students. To measure students' ability to reason scientifically, the use of laboratory reports assumes that classroom experiments and resulting write-ups mirror real scientific practice to an appropriate degree. Laboratory reports will be an effective source of data on students' scientific reasoning and inquiry skills if, and only if, students are allowed to engage in open-ended experimental situations which do not have pre-determined 'correct' answers. While the use of such inquiry-based labs is not yet universal (Basey, Mendelow, and Ramos 2000), they are commonly recognised as an instructional 'best practice' in higher education science classrooms (Boyer Commission 1998, 2001; Committee on Undergraduate Biology Education 2003; National Science Board 2010) and so their use is likely to increase.

To measure students' written scientific reasoning skills, a measurement tool was required, and a universal tool, one that is independent of subject matter or topic, would have greater utility than an assessment tool that was content- or topic-specific. Given that there are commonalities in scientific writing, in so much as most papers use similar category headings (Introduction, Materials and Methods, Results, etc.), it seemed plausible that a single rubric could measure aspects of critical thinking and scientific inquiry across a wide range of courses, topics and assignments. Additionally, students learn best when performance goals are made explicit (Aleven and Koedinger 2002; Campbell et al. 2000; Lin and Lehman 1999), so employing the same rubric across multiple courses (and making it clear to the students that the same rubric is being used) might improve student learning by defining and delineating a progression of learning goals and expectations (Schwarz et al. 2009). Such a rubric might thereby function as an instructional aid in and of itself.

We were also hopeful that such a rubric would facilitate consistent evaluation of lab reports by instructors while reducing the time required to design assignments and corresponding evaluation criteria. Science graduate students are often instructors in research-oriented institutions of higher education, but are traditionally given little pedagogical training, professional development opportunities or support for the development of their teaching abilities (Luft et al. 2004; Shannon, Twale, and Moore 1998). Additionally, their effectiveness as teachers or the reliability of their grading is virtually unstudied (but see Volkmann and Zgagacz 2004). We anticipated that a universal rubric would provide a useful tool to assist graduate students in becoming more efficient and more consistent evaluators of student work while also providing a glimpse into this common, but largely unstudied, instructional situation.

A universal rubric that could be applied across multiple courses within a curriculum could also allow assessment of the curriculum as a whole for departmental

accreditation. Application of the rubric to student work from freshman to senior year would provide a mechanism for determining the overall effectiveness of the curriculum in producing scientifically proficient students. Use of a standardised set of criteria across multiple levels of courses can align and propagate expectations of student performance (e.g. creates a 'learning progression', Schwarz et al. 2009). Application of a common metric to assignments across multiple courses can also highlight gaps between expectations and instruction, misalignments in expectations across course sequences and other scaffolding failures that may be impacting student achievement. Lastly, defining a common set of learning goals and performance expectations can facilitate the development of students' skills because even if a rubric is used only for summative assessment in a given course, using it repeatedly constitutes formative feedback over the span of a student's educational career (Black and Wiliam 1998; Yorke 2003).

A review of the literature revealed that no published rubrics addressed our specific needs. A number of rubrics were found which evaluate critical thinking in student writing, but none of them focused on the particulars of scientific reasoning or writing and none were intended for higher education. Several studies did describe relevant criteria for science writing at the university level (Haaga 1993; Topping et al. 2000), but lacked the description of performance levels which give a rubric its utility. In addition, the criteria were too assignment specific to be useful in other courses (Kelly and Takao 2002). Halonen et al. (2003) describe a comprehensive rubric for development of students' scientific reasoning skills, but it is not oriented toward written work, nor has it undergone reliability testing. Tariq et al. (1998) developed a rubric for assessing Honours projects produced by undergraduates in the biological sciences and the presentation portion of their rubric contained eight relevant criteria, but their analysis of the utility of the rubric was focused on comparing it with the incumbent assessment tool rather than determining its stand-alone reliability or validity. Once it was determined that there was a paucity of assessment tools that addressed university-level scientific reasoning and writing, we commenced development of an effective tool to fill this need.

Explicitly, our intent was to develop a tool which would: (1) improve the consistency of grades given by instructors (particularly graduate student instructors in large, multi-section introductory courses), (2) reduce the cognitive load on instructors for designing new assignments, (3) provide consistent meaningful educational goals across courses and make those learning goals explicit to our undergraduate students, and (4) allow comparison of student performance across courses and over time independent of assignment content.

Methodology

The development of the Rubric for Science Writing (Rubric) occurred over a three-year period and culminated in the validity and reliability studies reported here.

Rubric development and sources of validity

In the validation of an instrument, the criteria and content should be described and justified in terms of the construct the assessment is intended to represent (American Educational Research Association (AERA), American Psychological Association, (APA) and NCME 1999). Validity evidence based on the appropriateness of the

content (i.e. criteria in the Rubric) was derived from four sources: (1) relevant rubrics in the literature, (2) comparison to professional referee criteria, (3) consulting pedagogical experts, and (4) multiple rounds of recursive feedback from stakeholders who also served as content experts (e.g. Halonen et al. 2003; Sevian and Gonsalves 2008). Such recursive consultation with experts is a common method of constructing group consensus concerning ill-defined problems (Feldon 2007; Linstone and Turoff 1975) and provides a method for determining the desired performance of undergraduates in a skill domain as complex as scientific writing.

Review of relevant rubrics in the literature

Review of the relevant literature was initiated prior to the construction of the Rubric and continued throughout the process of rubric development. Criteria and rubrics found in the literature were combined with professional referee criteria to provide triangulation that our Rubric criteria represented aspects of the scientific process and scientific writing that were valued by the wider science education community as well as practicing scientists (Table 1). All sources were overlaid and the focus and intent of the resulting common criteria suggest that our Rubric captures core values. For example, the criterion that the 'introduction of the paper provides appropriate context for the work' was supported by all the relevant studies found in the literature. The majority of the criteria found support from three or more other studies and each criterion in Table 1 is supported by at least one other published study. Willison and O'Regan's (2007) insightful and relevant framework for assessing students' scientific skills was published after our Rubric and so constitutes a post-hoc test of the relevance of our criteria. Direct correspondence between the majority of criteria in our Rubric and theirs suggests that these are commonly agreed upon concerns and themes for assessing student scientific abilities (see Table 1).

Review of professional referee criteria

Criteria used by professional referees to review manuscripts were also compiled. The following relevant criteria consistently appeared: *significance of the work* and *quality of the methodology* (Cicchetti 1991; Marsh and Ball 1981, 1989; Petty, Fleming, and Fabrigar 1999), *writing quality* (Cicchetti 1991; Marsh and Ball 1981, 1989), *literature review* (Cicchetti 1991; Marsh and Ball 1981, 1989), *succinctness* (Cicchetti 1991), *originality* (Cicchetti 1991) and *theoretical context* (Petty, Fleming, and Fabrigar 1999). Thus, there appears to be a support within both the scientific community and the educational community that *methodological competency*, *appropriate context*, *adequate literature review*, *scientific merit/significance* and *writing quality* are fundamental attributes of high quality scientific reasoning and science writing.

Validity derived from content experts and stakeholders

Ultimately, the validity of the Rubric as a tool for assessing scientific writing is fundamentally derived from the perceptions of stakeholders and potential users (e.g. faculty and graduate teaching assistants) (Dalkey 1969). If stakeholders have doubts about an instrument, it can often have significant negative consequences on adoption of a new assessment tool (Tariq et al. 1998). Validity also hinges on the intended use of the

Table 1. Criteria for evaluating scientific reasoning in higher education.

	Rubric for Science Writing (this study)	Haaga (1993)	Kelly and Takao (2002)	Halonen et al. (2003)	Tariq et al. (1998)	Topping et al. (2000)	Willison and O'Regan (2007)[a]
Study context	Undergraduate biology laboratory reports	Graduate psychology manuscripts	Undergraduate oceanography scientific report	Higher education curriculum outcomes in psychology	Undergraduate biology and biomedical sciences honours projects	Graduate psychology term papers	Undergraduate mentored independent research
Instrument reliability	$g = 0.85$	$r = 0.55$	Not tested	Not tested	Not tested – only compared new and old schema	Not tested	Not tested
Performance levels	Four levels: 'Not addressed' to 'Proficient'	None specified	None specified	'Before training' to 'Professional' (Five levels)	Five levels indicating letter grades	None specified	Five levels based on degree of student intellectual autonomy
Criteria							
	Introduction: Context of study	Background (primary lit) is covered adequately	Clear distinction between portions of the theoretical model supported by data/ background knowledge and those which are still [untested]	Degree of theoretical/ conceptual framework (Table 2, p. 199)	Comprehension of concepts and aims/ context; depth of understanding; appreciation of relevance of findings to subject area	Clear conceptualisation of the main issues Literature review	Students embark on inquiry and to determine a need for knowledge/ understanding
	Hypotheses are testable		A clear, solvable problem is posed …				
	Hypotheses have scientific merit		… based on an accurate understanding of the underlying theory				The most autonomous level expects students to 'generate questions/aims/ hypotheses based on experience, expertise and literature'

Table 1. *(Continued).*

	Rubric for Science Writing (this study)	Haaga (1993)	Kelly and Takao (2002)	Halonen et al. (2003)	Tariq et al. (1998)	Topping et al. (2000)	Willison and O'Regan (2007)[a]
	Methods: Experimental design		Multiple kinds of data are used when available	Sophisticated observational techniques, high standards for adherence to scientific method, optimal use of measurement strategies, innovative use of methods (Tables 1 and 3, p. 198–9)	Appropriateness [and excellence] of methods and/or experimental design		Students find/generate needed informational data using appropriate methodology
	Results: Data selection		Available data are used effectively; data are relevant to the investigation			New data (type, range, quality)	
	Results: Data presentation		Observations are clearly supported by figures				
	Statistics			Uses statistical reasoning routinely (Table 3, p. 199)	Findings handled competently in a scientific way, appropriate analyses as applicable		Students organise information collected/generated

Table 1. *(Continued).*

Rubric for Science Writing (this study)	Haaga (1993)	Kelly and Takao (2002)	Halonen et al. (2003)	Tariq et al. (1998)	Topping et al. (2000)	Willison and O'Regan (2007)[a]
Conclusions based on data	Conclusions follow logically from evidence and arguments presented	Conclusions are supported by the data; text clearly explains how the data support the interpretations	Uses scepticism routinely as an evaluative tool Seeks parsimony (Table 5, p. 200)	Conclusions are comprehensive and in agreement with findings Interpretation of findings is appropriate in all or most instances	Conclusions/ synthesis	Students synthesise and analyse new knowledge
Limitations/ significance/ future directions			Understands limitations of methods, 'bias detection and management' (Table 3, p. 199)	Includes many excellent ideas for continuation		Students critically evaluate information/data and the processes to find/ generate this information/data
Primary literature		Data adequately referenced	Selects relevant, current, high quality evidence, uses APA format (Table 6, p. 201)	Literature review is appropriate and used [primary] material	References Literature review	The most autonomous level expects students to place their work within the context of the primary literature
Writing quality	Well-written (clear, concise, logical organisation and smooth transitions)	Clear, readable focused and interesting; accurate punctuation and spelling; technical paper format (complete and correct)	Organisation, awareness of audience, persuasiveness, grammar (Table 6, p. 201)	Material is well-organised in logical sequence suiting specific nature of material; presentation of material is clear, easy to follow and shows imagination	Structure (headings, paragraphs); precision and economy of language; spelling, punctuation syntax	Students communicate knowledge and understanding and the processes used to generate them

Table 1. *(Continued).*

	Rubric for Science Writing (this study)	Haaga (1993)	Kelly and Takao (2002)	Halonen et al. (2003)	Tariq et al. (1998)	Topping et al. (2000)	Willison and O'Regan (2007)[a]
Additional criteria		Goals of the paper are made clear early Scope of the paper is appropriate (not over-reaching or over broad)	Clear distinction between observations and interpretations Epistemic level: Arguments build from concrete data to more abstract theory; each theoretical claim supported by multiple data sources	Awareness, evaluation of and adherence to ethical standards, and practice (Table 4, p. 200) Scientific attitudes and values: enthusiasm, objectivity, parsimony, scepticism, tolerance of ambiguity (Table 5, p. 200)	Ability to problem solve; review and modification of plan; motivation; competence and independence; plan of action; organisational skills	Psychology content Advance organisers (abstract, contents) Originality of thought Action orientation	

[a]Willison and O'Regan (2007) was published after our Rubric was developed.
Note: If not indicated directly in the table, quotations were found as follows: Kelly and Takao (2002), Table 1, p. 319; Haaga (1993), Table 1, p. 29; Tariq et al. (1998), Figure 5, p. 228; Topping et al. (2000), Appendix 1, p. 167; Willison and O'Regan (2007), Figure 1, p. 402–3.

tool. Assessment tools should be evaluated in terms of the intended purpose of the tool and the consequences if a tool provided invalid information (Moskal and Leydens 2000). Thus, as this instrument is intended to aid instruction, the vast proportion of its validity derives from whether or not the intended end users believe the Rubric to be appropriate in its priorities and scope (e.g. content validity – does it ask the right questions?) and whether or not the scores generated by the Rubric appear to provide meaningful information about students' ability to achieve those desired outcomes (Do the scores match the perceived quality of the work?). As marks for coursework ultimately come from the opinion of the instructor, if they perceive the Rubric as providing valid and meaningful information, then it has equal validity with whatever assessment method was used in the absence of the Rubric.

Our intent for this Rubric was for it to be applicable in assessing scientific writing across multiple courses; so the criteria in the Rubric were derived from the departmental curriculum goals, informal discussion with select faculty and review of the literature, all of which were synthesised by the lead author who then designed and tested the first iteration of the Rubric over two semesters in her introductory biology laboratory sections. When Rubric constructs were consistently effective in assessing meaningful aspects of her students' ability to engage in scientific reasoning and writing, review by additional faculty in the department commenced. The development of the Rubric therefore focused on recursive evaluation and revision by the intended end users, who were also experts in the desired performance domain (e.g. tenure-track faculty, instructional faculty and graduate teaching assistants in the biology department), with significant contributions from pedagogical experts.

This methodology is a modified Delphi technique wherein expert opinions are synthesised in order to characterise complex processes or gather collective subjective judgements (Linstone and Turoff 1975). A modern version of this methodology is known as 'cognitive task analysis' (Feldon 2007). In short, both techniques use recursive surveying of panels of experts, coupled with increasing synthesis of information and feedback, to generate a consensus. We used recursive one-on-one interviews of anonymous experts (in relation to other experts on the panel) and controlled feedback (summarised and stripped of opinion or other inflammatory content) to encourage free communication as recommended by Dalkey (1969). Particular attention was also paid to the selection of experts to ensure a broad representation of perspectives (Reeves and Jauch 1978). One limitation of using a consensus derived in this way is that it may represent a compromise solution rather than an optimum solution, but when success depends upon a sense of ownership by constituents (as in the case of curriculum or assessment design), consensus by compromise is appropriate and perhaps even necessary (Reeves and Jauch 1978).

In the second year of Rubric development, the instrument was incorporated into introductory biology courses, and two permanent full-time instructors as well as 10-tenured or tenure-track faculty in the biology department reviewed the Rubric criteria and performance levels and provided feedback as to its appropriateness for assessing students' scientific writing in courses ranging from freshman level to upper division. Selected faculty were practicing scientists from a variety of subfields within biology (molecular biology, developmental biology, genetics, evolution, ecology) who all possessed active external research funding. Additionally all reviewing faculty were active instructors and two had won university-wide teaching awards.

Reviewers were asked to focus their critique on whether or not the criteria represented valid and meaningful educational goals for scientific reasoning and writing,

and whether or not the performance levels described common student performances and desired outcomes. Additionally, all instructors associated with the introductory biology courses in which the Rubric was being used provided ongoing feedback and suggestions for revisions. Additionally, a practicing scientist and award-winning instructor from an external department was solicited both to determine the potential utility of the Rubric outside of biology and as a brief check for any bias due to departmental culture or practice (Dalkey 1969).

In the third year of the development process, the validity associated with the Rubric was investigated through feedback from biology graduate teaching assistants (n = 19) whose content expertise spanned the spectrum of the biology department, including bioinformatics, cell and molecular biology, genetics, plant biology, ecology, and evolution. The graduate teaching assistants were given a diverse set of student papers from multiple biology courses and asked to score those papers using the Rubric in a single supervised scoring session. Graduate students were explicitly asked if the resulting scores seemed indicative of overall paper quality (i.e. scientific writing) and whether or not the Rubric assessed meaningful and informative aspects of student performance. We also asked for feedback about the ease of use and pragmatic effectiveness. All written comments, both short answers to specific questions and notes written on the rubrics or papers themselves, were collected and reviewed by the lead author. Conflicting comments or comments that suggested a need for revision in either content or focus of Rubric elements were followed up with one-on-one interviews between the lead author and reviewers. All feedback was therefore eventually synthesised and incorporated into the finalised Rubric. Table 2 lists the finalised Rubric criteria. A complete copy of the finalised Rubric is attached as Appendix 1.

Lastly, the Rubric was reviewed by four faculty members with doctorates in educational research, two within our institution and two at other institutions. Reviewers were asked to critique the Rubric in terms of construct validity (Moskal and Leydens 2000) and pedagogical appropriateness.

In sum, all faculty involved in the review process reported that they found the revised criteria and performance levels appropriate in scope, focus and level for the stated undergraduate instructional aims. Graduate teaching assistant feedback indicated that the Rubric provided meaningful measures of student performance, i.e. scores generated by the Rubric aligned with the sense of the overall quality of the paper and appropriately identified areas of strength and weakness.

Testing the reliability of the Rubric in realistic conditions

Source of student papers

To test the reliability and universality of the Rubric, student papers were selected from three different university biology laboratory courses to represent a spectrum of content areas and student performance levels (Table 3). These courses included the first and second semesters of the introductory biology course sequence for majors (BIOL 101 and 102) and the laboratory associated with a required majors course (BIOL 301) intended to be taken by sophomores, but often populated by juniors and seniors. In each course, students engaged in an open-ended investigation and wrote a laboratory report detailing their work and findings (Table 3). From the several hundred students populating these courses in a single semester, a subset of 45 to 50 papers were selected from each course based on the following criteria: (1) paper and graphs were complete, on topic and without plagiarism; (2) paper was authored by a biology major currently

Table 2. List of criteria and definitions in the Rubric for Science Writing.

Criteria	Definition
Introduction	
Context	Demonstrates a clear understanding of the big picture; Why is this question important/interesting in the field of biology?
Accuracy	Content knowledge is accurate, relevant and provides appropriate background including defining critical terms.
Hypotheses	
Testable	Hypotheses are clearly stated, testable and consider plausible alternative explanations.
Scientific merit	Hypotheses have scientific merit.
Methods	
Controls and replication	Appropriate controls (including appropriate replication) are present and explained.
Experimental design	Experimental design is likely to produce salient and fruitful results (actually tests the hypotheses posed).
Results	
Data selection	Data chosen are comprehensive, accurate and relevant.
Data presentation	Data are summarised in a logical format. Table or graph types are appropriate. Data are properly labelled including units. Graph axes are appropriately labelled and scaled and captions are informative and complete.
Statistical analysis	Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained.
Discussion	
Conclusions based on data selected	Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.
Alternative explanations	Alternative explanations (hypotheses) are considered and clearly eliminated by data in a persuasive discussion.
Limitations of design	Limitations of the data and/or experimental design and corresponding implications for data interpretation are discussed.
Significance of research	Paper gives a clear indication of the significance and direction of the research in the future.
Primary literature	Writer provides a relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided) using primary literature.
Writing quality	Grammar, word usage and organisation facilitate the reader's understanding of the paper.

Note: Primary literature is defined as published paper which have been peer reviewed, report original data (not a review), the authors collected the data and a non-commercial scientific association publishes the journal. The full Rubric is attached as Appendix 1. A copy of the Scoring Guide version (rubric plus examples of student performance at each level) is also available from the corresponding author on request.

Table 3. Description of biology courses and assignments which generated writing samples used in this study.

Course no. (Level)	Course enrolment	Content area: Description of experimental investigation	No. of papers selected
101 (1st semester biology majors)	403	*Genetics*: Determine the Mendelian inheritance pattern of an atypical phenotypic trait in fruit flies (*Drosophila melanogaster*) by performing live crosses and analysing the results.	49
102 (2nd semester biology majors)	328	*Evolution*: Determine whether or not evolution occurred in a population of Galapagos Finches using a multi-year dataset (*Galapagos Finches* www.iqwst.northwestern.edu/finchesdownload.html).	45
301L (lab for required major course)	135	*Ecology*: Determine if the abundance and distribution pattern of dandelions in a field correlates with environmental factors such as sun and shade.	48

Note: All investigations culminated in students writing laboratory reports.

enrolled in the biology programme; (3) no more than five papers were selected from any one lab section (maximum enrolment of 24 students per section, 33 sections sampled); and (4) within each laboratory section at least one paper was selected from a student who earned an 'A' in the course and at least one from a student who earned a 'D'. Efforts were made to select papers representing the available spectrum of quality (as determined by course grades) for each lab section sampled. Selected papers were then stripped of all author-identifying information, assigned an ID code and standardised for font, margins and line spacing before being printed and copied.

Graduate student raters: selection and apportionment

Two sets of graduate students were recruited from within the biology department. The biology graduate programme at our institution is a large (approximately 70 graduate students) primarily doctoral programme mentored by approximately 30 faculty with notable levels of externally funded research and is therefore representative of many science departments at large, research-oriented institutions. The first set of graduate students ($n = 9$) received several hours of training on how to use the Rubric and scored papers under controlled conditions within 48 hours (hereafter referred to as 'trained raters'). The second set of graduate students ($n = 8$; hereafter referred to as 'natural raters') received approximately 10 minutes of verbal instructions and a handout with content matching that of Table 2.

Raters were assigned among the three sets of papers (three raters per paper) according to their years of overall teaching experience, teaching experience in that particular course as well as area of content expertise (Table 4). Each group of three raters included at least one person with less than three semesters of teaching experience and at least one person with more than three semesters of experience (usually seven or more), as prior teaching experience could potentially affect ability and/or comfort in scoring student papers. These groups also represented typical laboratory teaching conditions at our

Table 4. Prior experience of graduate student raters.

Course	Rater type	No. of semesters of teaching experience per rater	Raters have taught this assignment?
101	Trained	2,2,5	Y,Y,Y
	Natural	1,2,4	Y,Y,Y
102	Trained	3,3,11	Y,Y,N
	Natural	1,4,4	Y,Y,Y
301	Trained	3,7,7	Y,N,Y
	Natural	3,7	Y,Y

Note: All raters had strong content background in the subject matter relevant for their assigned papers. Codes are as follows: (Y) yes, (N) no. Data are organised respectively meaning that the first code in each column refers to the same individual (e.g. in Row 1 the first two raters each had two semesters of teaching experience and had taught BIOL 101 including this same assignment in the past while the third rater had five semesters of teaching experience and had also taught this assignment in the past). The average number of semesters of teaching experience for trained raters was 4.8 + 3.0 and for natural raters was 3.4 + 1.9.

institution. The BIOL 301 group contained only two raters because the pool of graduate students with experience in teaching in that subject area was more limited; one graduate student had three semesters of experience and the other had seven semesters.

Rater training

Trained raters received five hours of training led by the third author. Training included an overview of the project, discussion of sources of bias in scoring, and explanation of the Rubric and associated scoring guide (version of the Rubric which included examples of student work at each performance level for each criterion, see Table 5 for an example). Raters individually scored three exemplar papers representing poor, average and excellent examples of student papers from the course to which they had been assigned. Each group of raters then discussed any discrepancies in the scores assigned to the exemplars until consensus was reached for all scores and all papers (Johnson et al. 2005). This discussion thereby served to calibrate each group of raters and ensure that each criterion was interpreted in a consistent manner. In addition, raters received training in the use of augmentation. Raters first assigned a score and then could augment that score with a plus or minus to indicate that the response is slightly higher than a typical paper at that level or slightly lower than a typical paper at that performance level (Johnson, Penny, and Gordon 2000; Penny, Johnson, and Gordon 2000a) (Table 5). Augmentation allows for finer resolution among performance levels (Penny, Johnson, and Gordon 2000b). Scores and augmentations were translated into numerical equivalents. If the performance reflected a typical response, then the non-augmented (i.e. integer) score was reported.

Natural rater conditions

Natural raters received brief verbal instructions on the purpose of the assignment and a handout listing the Rubric criteria with definitions (e.g. Table 2). Natural raters used the same point scheme for scoring as did trained raters, but had a week to score the papers and were unsupervised as they did so. This is representative of the degree of guidance commonly provided by faculty to graduate teaching assistants at our institution.

Table 5. Example of a page from the Rubric for Science Writing scoring guide that was used during the rater training.

Criteria	0 Not addressed 0+	1– Novice 1+	2– Intermediate 2+	3– Expert 3+
Introduction: Context				
Demonstrates a clear understanding of the big picture.	The importance of the question is not addressed.	The writer provides a generic or vague rationale for the importance of the question.	The writer provides one explanation for why others would find the topic interesting.	A clear sense of why this knowledge may be of interest to a broad audience.
Why is this question Important/ interesting in the field of biology?	How the question relates within the broader context of biology is not addressed.	The writer provides vague or generic references to the broader context of biology.	The writer provides some relevant context for the research question(s).	Writer provides explanation of gaps in understanding and how this research will help fill those gaps.

Note: The scoring guide includes examples of student work at each performance level to aid raters in interpreting criteria. The top row indicates each performance level and its associated point values for the stated criterion. The next two rows are the title of a criterion, its definition and a description of student performance at each level. Examples of student work for each level of proficiency are given below:

Example of a novice (1–) response:
'Gene sequencing is done by first PCR/ing the DNA and then cutting it with primers to determine the sequence of nucleotides. Many scientists like to know the sequence of genes so that they can perform other experiments on those organisms' ('so they can perform other experiments' is too vague – no sense of context is provided').

Example of a novice (1) response:
'Plant competition is a frequent ecological interaction'. [Definition of intra vs. inter specific competition provided.] 'There are several factors that influence competition between plant species: density of plants, and amount of available nutrients and sunlight … Tilman proposes that the outcome of competition can either be at equilibrium, or can oscillate overtime' (*While a reference to someone else's work is provided, there is no sense of why competition is worth studying or what the consequences are of various competitive outcomes (no sense of importance or broader context))*.

Example of an intermediate (2) response:
'One of biology's major ideas relates to the question of why do offspring resemble their parents'. [Historical context is then followed by definition of phenotype and Mendel's three laws and a rationale for why the fruit fly is used as a model organism for genetics.] 'The purpose of this experiment is to determine whether sex-linked traits are identifiable by the ratios of phenotypes in offspring. This experiment is important because by examining and analysing inheritance patterns (i.e. using Punnett squares), scientists can determine the probability of inheriting certain traits' (*implicit sense of importance, historical context and justification for organism provided*).

Example of an expert (3) response:
'The effect of spore dispersal distance on fungal biodiversity is a subject that is superficially well-understood by scientists; however, much of the intricate workings of genetic variation in individuals and in the resulting populations are yet to be discerned. Fungi have the ability to reproduce both sexually and asexually, but the relative impact of these propagation methods on their genome is unknown and paramount to comprehending their evolutionary and reproductive history. Such results would enlighten our understanding of the co-evolution of plants and fungi as well as provide a foundation for assessing the health of populations and fungal biodiversity' (*broad context is provided as well as identification of gaps in knowledge and why they matter*).

Scoring of student papers

All raters individually rated the complete set of student papers for their assigned course (Table 3) using the same point scheme. Each criterion was worth a maximum of 3 points. Raters matched the student performance for each criterion to the appropriate performance level and assigned the relevant point value (e.g. Table 5) by recording the points earned by each paper for each criterion on a spreadsheet. Both trained and natural raters were allowed to augment scores with '+' or '–' as described above. There was no discussion among raters during the scoring phase. Scores assigned by the different raters for the same paper were then compared by generalisability analysis to calculate the reliability of the Rubric criteria and the cumulative total score per paper.

Relevant institutional context

It should be noted that at our institution, all graduate students must teach introductory biology before moving on to more autonomous teaching roles in other courses. Introductory biology is our most pedagogically supported course. Teaching assistants are required to attend a weekly meeting typically lasting two to three hours during which they receive pedagogical and logistical support and training for that week's teaching duties. Faculty laboratory coordinators monitor grade distributions and meet with teaching assistants with unusually high or low class averages, and graduate teaching assistants are exposed to the Rubric criteria whenever those criteria are incorporated as part of the course assignments.

The Rubric described by this paper was incorporated into the introductory biology courses for several years prior to the collection of these data. Thus, both our natural raters and our trained raters had prior experience with the Rubric and were likely at least somewhat familiar with the criteria and its application to lab reports. Nevertheless, these 'natural' raters represent the baseline condition of graduate students as instructors in our department. Readers are cautioned that it is likely that we found smaller differences between trained and natural raters than might be found in other institutions as a result of the pedagogical training provided in introductory biology at our institution.

Data analysis

Rubric reliability data were analysed using generalisability (*g*) analysis (Crick and Brennan 1984) which determines the portion of the variation in scores which is attributable to actual differences in the quality of the papers rather than variation among raters or assignments or variation due to interaction among factors such as student-assignment or rater-assignment interactions (Brennan 1992; Shavelson and Webb 1991). For example, a generalisability score of 1.0 means that *all* the variation in scores is due to differences in quality among the student papers and that no error was introduced (all raters were perfectly consistent regardless of student, assignment, etc.). In contrast, a generalisability score of 0.0 means that *none* of the variation in scores among papers was attributable to actual differences in quality (all the variation was entirely due to other sources such as rater inconsistency instead). All generalisability scores reported here were generated by comparing scores from three trained raters unless otherwise specified.

Results and discussion

Reliability of cumulative scores

Reliability of the Rubric using the cumulative (total) score for each paper was high ($g = 0.85$) for each of the three different classes. In general terms, these results mean that 85% of the variation in the total scores was reflective of actual differences in the quality of the papers (rather than rater inconsistency or other sources of error), and that scores generated with the Rubric under similar circumstances would produce highly reliable grades. These results are comparable with other published educational rubrics for writing (Table 6). Additionally, while there are some reasons to expect that professional peer reviews would show greater disparity in scores, the average reliabilities reported for 16 studies on the reliability of professional peer review were notably lower than that returned by our Rubric (Table 6).

The fact that the Rubric returned high reliability values for each set of papers from three different courses suggests that the criteria tested here are assessable across subfields of biology. The maximum reliability scores for individual criteria are evenly distributed among the courses (101 and 102 each have 4 maxima, 301 has 6 maxima; see bold values in Table 7). Thus, the Rubric as a whole appears reliable across a variety of subject areas within biology. Informal discussion among science faculty outside of biology has suggested that the Rubric may also be applicable to

Table 6. Comparison of the reliability of the Rubric for Science Writing with professional peer review and relevant published rubrics for evaluating student writing.

Citation	Reliability statistic	No. of criteria	No. of raters	Reliability value
Rubric for Science Writing (this study)	g	15	3 1[a]	0.85 0.65, 0.66
Studies of rubrics in educational settings				
Baker et al. (1995)[b]	α	6	4	0.84–0.91
Cho et al. (2006)[b]	α	3	5	0.88[c]
Haaga (1993)[d]	r	4	2	0.55
Marcoulides and Simkin (1995)[2]	g	10	3	0.65–0.75
Novak et al. (1996)[b,e]	g	6	1[a], 2	0.6, 0.75
Penny, Johnson, and Gordon (2000)[b]	phi	6	2	0.6–0.69
Studies of professional peer review				
Cicchetti (1991)[f]	r	Various	1[a]	0.19–0.54 (median 0.30)
Marsh and Ball (1989)[f]	r	Various	1[a]	0.27 ± 0.12 (ave ± SD)
Marsh and Ball (1981)	r	5	2	0.51
Marsh and Ball (1989)	r	4	1[a]	0.30
Marsh and Bazeley (1999)	phi	Holistic	4	0.704

[a]Single rater reliabilities were calculated from two-rater or three-rater data. [b]Non-scientific writing samples. [c]Reliability produced by undergraduate peers rather than trained raters. [d]List of criteria only, not a rubric. [e]Multiple rubrics reported in this study; these results refer to the WWYR rubric. [f]Meta-analyses of multiple studies. Note: Professional peer review employs lists of criteria rather than rubrics with defined performance levels that may account for some difference in reliability scores. All reliability values are quoted directly from cited papers.

Table 7. Reliability of individual Rubric criteria using trained raters.

	Biology course			
	101	102	301	Average across all courses
Criteria	$n = 49$	$n = 45$	$n = 48$	$n = 142$
Introduction				
Context	0.67	**0.83**	0.50	0.67
Accuracy and relevance	**0.67**	0.47	0.65	0.60
Hypotheses				
Testable	0.70	0.70	**0.81**	0.74
Scientific merit	**0.76**	0.66	0.67	0.70
Methods				
Controls	—[a]	0.00[b]	0.16[b]	n/a
Experimental design	0.20	**0.89**	0.57	0.55
Results				
Data selection	0.50	0.53	**0.66**	0.56
Data presentation	**0.77**	0.72	0.64	0.71
Statistics	0.59	0.02[2]	**0.62**	0.61
Discussion				
Conclusions based on data	0.63	0.60	**0.65**	0.63
Alternative explanations refuted	**0.73**	0.55	0.72	0.67
Limitations	0.57	**0.83**	0.60	0.67
Significance	0.56	**0.81**	0.79	0.72
Primary literature	0.57	0.85	**0.94**	0.79
Writing quality	0.42	0.35	**0.71**	0.49
Total score	**0.85**	**0.85**	**0.85**	0.85

[a]The trained raters for BIOL 101 did not perceive the genetics assignment as providing a traditional control and chose to not rate this criterion. Natural raters scoring 101 papers achieved a three-rater reliability of 0.74 for this criterion however. [b]See text section on low-criteria reliabilities for explanation.
Note: Reliability values were calculated using generalisability analysis (g). Values in bold are the maximum reliability score per criterion. Sample sizes reflect the number of unique papers scored per course. All values reported are three-rater reliabilities using trained raters.

other non-biological scientific fields as well. Testing the Rubric in novel contexts and across scientific, mathematic and engineering disciplines would be a fruitful line of future research that would highlight the degree and nature of conceptual gaps and bridges between these commonly aggregated fields.

Reliability of individual criteria

The reliability of individual criteria was lower than that of the cumulative (total) score for a paper (Table 7). This is analogous to the reliability and informational value of a single exam question versus the total exam score. With a few exceptions, most criteria were found to be reliable in a variety of contexts. The minimum three-rater reliability (g) across all three datasets was 0.20 and the maximum was 0.94, with

an average reliability of any single criterion being 0.65. This result excluded the criteria of *Methods: Controls* for all three datasets and *Results: Statistics* for BIOL 102 and *Methods: Experimental Design* for BIOL 101 which will be discussed separately (refer to Table 2 or Appendix 1 for full descriptions of criteria). Maximum reliabilities for any single criterion ranged from 0.62 to 0.94. These results indicate that each criterion is reliable in at least a subset of situations. Additionally, other published rubrics have reported individual criterion reliabilities as low as $g = 0.151$ for four raters, but those individual criteria aggregate into a total score reliability of $g = 0.53$ (Baker et al. 1995). This pattern of overall or total scores having equal or higher reliabilities than criterion scores was also found in other studies of student work (Haaga 1993; Klein et al. 1998) and professional peer review of journal submissions and grant proposals (Cicchetti 1991; Marsh and Ball 1981; Marsh and Bazeley 1999). Therefore, instructors should not focus on a single criterion and assignments should incorporate multiple criteria in order for a total score to reliably reflect the quality of student work.

Failure to include some Rubric criteria in course assignment affects criterion reliability

The few criteria with low-reliability scores mentioned above (criteria reported in Table 7 with reliabilities at or near zero) may have been due to the absence of Rubric criteria into the course assignments rather than an issue with the criteria themselves (Table 8). There was a noticeable relationship between the degree to which Rubric criteria were included in the course assignment and the reliability of scores generated for that criterion (Table 9).

Exclusion of criteria from the assignment likely affects reliability scores because generalisability analysis assumes that there will be variation in performance among papers and perceives such a lack of variation as an indicator of low-criterion reliability. But, if students fail to even attempt a criterion (such as the use of controls in experimental design or the use of statistics), the lack of variation is an accurate reflection of performance. For example, *Methods: Control* was a criterion that was clearly omitted from the BIOL 102 assignment, and 132 of 135 scores earned for this criterion were zeros. Similarly, the same criterion was omitted from BIOL 301, and 126 of the 144 scores were zeros. In BIOL 102, the instructions given to students did not explicitly ask students to incorporate controls or use statistics and so most students did not attempt any statistical comparisons, and 123 of the 135 scores given by raters were zeros. As further support for the conclusion that those low reliabilities are not meaningful, these criteria all performed well in at least one of the three courses. Except for *Methods: Controls* (which was not incorporated into any assignment), each of the other criteria had a reliability at or above 0.62 (Table 7) in at least one course. The low-reliability scores for these criteria are therefore interpreted as an artefact of the assumptions behind generalisability analysis.

It should be noted that the criterion *Methods: Controls* appears to remain essentially untested. It should also be noted that one criterion (*Hypotheses have scientific merit*) does appear to be a possible exception to this pattern. It was not included in any assignment, yet raters produced highly reliable scores. It is perhaps not really an exception, however, in that all assignments did require students to *pose hypotheses*. The high reliabilities are likely because student hypotheses

Table 8. Inclusion of Rubric criteria into course assignments.

	Criterion incorporated into the course assignment?		
Rubric criteria	101	102	301
Results: Statistics	Yes		Yes
Discussion: Conclusions based on data selected	Yes	Yes	Yes, but highly implicit
Hypotheses: Scientific merit			
Hypotheses: Testable and consider alternatives	Partial: 'clear with rationale'	Yes, verbatim	Yes, clearly stated
Results: Data presentation	Yes	Yes	Yes
Results: Data selection	Determined by instructor	Yes	Yes, but implicit
Discussion: Alternative explanations		Yes	
Introduction: Accuracy and relevance	Yes	Yes, verbatim	
Discussion: Significance of research	Yes	Yes	
Discussion: Limitations of design	Yes		
Methods: Controls	Determined by instructor		
Methods: Experimental design	Determined by instructor	Yes, but implicit	
Introduction: Context	Yes	Yes, verbatim	
Writing quality	Yes	Yes	Yes
Primary literature	Yes, two required	Bonus only	Partially, primary literature optional

Note: Criteria are rank-ordered from least variable to most variable based on spread between minimum and maximum reliability per course. Blank cells indicate that the criterion was not explicitly mentioned in the assignment. For BIOL 101 and 102, alignment designations were derived directly from the grading rubric handed out to students in the class. For BIOL 301, no written assignment was given to the students, so alignment of the assignment with the rubric was determined by communication with teaching assistants.

Table 9. Summary of correspondence between the inclusion of criteria in an assignment and criterion reliability.

	Degree to which criteria were included in assignment		
	Explicitly	Implicitly	Absent
Average reliability score (g)	0.63	0.68	0.55
Standard deviation	0.12	0.25	0.27
n =	23	7	14

Note: Reliability scores of individual criteria in each course were categorised according to the degree of inclusion in that assignment (Table 7). Sample size is number of average reliability scores in that category (i.e. values from Table 6).

naturally vary in their quality regardless of whether or not instructors discuss this aspect of science. In sum, all other criteria performed reasonably whenever they were included in the assignment in a manner that allowed student performance to vary with ability.

Effect of omitting criteria from the assignment on student skill development over time

As previously described, whether or not criteria are included in the assignment instructions affects interpretation of scores. When we narrow our lens to only those criteria that were included in both the 101 and 102 assignments (Figure 1), we see a definitive increase in the average student score for most criteria. For example, 'hypotheses are testable and have scientific merit', 'data selection', 'conclusions are based on findings', 'discussion of limitations and overall writing quality all improve'. But no change is seen in 'introduction: context', and 'primary literature' actually declines. Review of the assignments (Table 8) reveals that the use of primary literature was optional in the BIOL 102 assignment whereas it was mandatory in the 101 assignment. The drop in average score for the use of primary literature in BIOL 102 thus likely represents a lesser effort on the part of the students rather than a decline in actual ability. Supporting evidence for this is that 62% of the students did not address primary literature at all in their BIOL 102 papers (score of zero for this criterion). If zero values are ignored, the average score for this criterion is 0.5 ± 0.2 which is similar to the value for BIOL 101 (0.6 ± 0.4). In sum, criteria that were actively included in both assignments either returned average scores that remained steady (two of nine) or improved from one course to the next (seven of nine).

It should be noted that in absolute terms, average criterion scores for 101 and 102 hovered firmly within the 'novice' performance level. This is appropriate as the performance levels were intended to measure change across multiple years and higher average scores would be expected from students in upper division courses.

Scores from the BIOL 301 class were similar to those from BIOL 101 and 102 (average criterion score = 0.86 ± 0.23 SD, still firmly within the novice performance level), but were not included in Figure 1 because post-hoc analysis of students

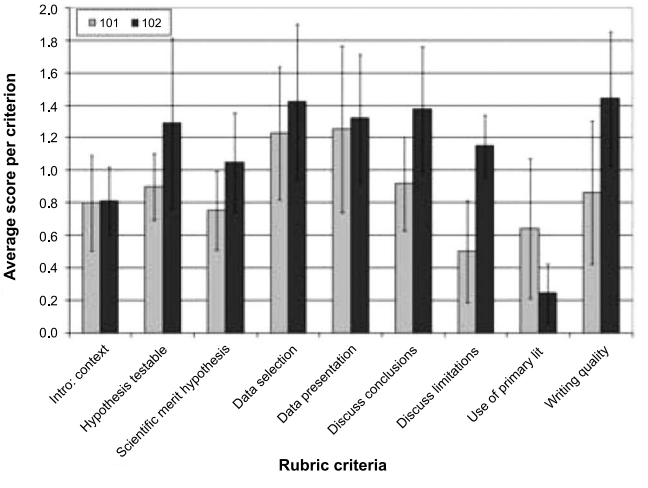

Figure 1. Student performance over time as assessed by the Rubric for Science Writing. Note: Only criteria which were included in both course assignments are reported (see Table 7). The average criterion scores were calculated by first averaging the three scores assigned by the raters for each criterion for any given student, then averaging across all students (BIOL 101, $n = 49$; BIOL 102, $n = 45$) within each course (Johnson et al. 2003).

enrolled in the BIOL 301 course revealed that many were transfer students from the local two-year college. Thus, these students were not a useful sample for determining the effect of the biology department curriculum on skill development as many of them had not completed any more of our biology courses than the students enrolled in the introductory sequence.

The Rubric was provided continually to the undergraduate population on the departmental website as well as being included to varying degrees with each assignment. It was the authors' intent that the Rubric should be an educative tool and function as a formative assessment mechanism as well as a means of improving the consistency of final marks. Specifically, student papers in BIOL 101 and 102 underwent formative peer review using Rubric criteria as an intentional component of the learning experience. Scored papers thus already incorporated changes made by student writers as a result of this peer feedback. A majority of undergraduate students who responded to an anonymous online survey (n = 1026 students) reported that the feedback provided by peer reviewers applying the Rubric criteria was helpful and improved the quality of their papers and that the act of reviewing itself was educationally beneficial (data to be reported in future papers).

Use of the Rubric as a programmatic assessment tool

Prior to the development of the Rubric, the biological sciences department was under the impression that the expectations for students had a consistent progression. After numerous curriculum committee meetings concerned with the scope and sequence of the biology courses, it became evident that some subsequent course assignments had omitted core expectations from previous courses. For example, students were required to do statistical analyses in BIOL 101 and reference primary literature, but those expectations were dropped in BIOL 102. Interviews with both faculty lab coordinators for the introductory sequence revealed that these gaps were unintentional oversights and that the relevant faculty were not aware of the gap until the Rubric was used in multiple courses. Another function of the Rubric, therefore, appears to be to identify curricular weaknesses or misalignments (Halonen et al. 2003). This is not a function specific to our Rubric; the application of any universal metric which has been aligned with curriculum goals would serve the purpose of assessing curriculum alignment and progression. Similarly, these results show that data on student achievement must be interpreted within the context of alignment between assignment goals and Rubric criteria. In short, if assignments do not ask students to perform the desired skills, students are unlikely to perform them well (if at all) and if repeatedly omitted from assignments, those skills are unlikely to improve over time.

Impact of the Rubric on graduate students' professional development as instructors

Besides its utility as a measurement instrument, the Rubric has potential to benefit instructors in the classroom. Science graduate students in particular often receive little support for their teaching and little training on pedagogical issues such as grading (Boyer Commission 2001; Davis and Fiske 2001; Luft et al. 2004). Use of a standardised rubric would provide consistency of expectations for students across multiple courses within a curriculum as well as save graduate teaching assistants from the work of developing their own grading schema. Given the general lack of attention to graduate students as instructors (e.g. Luft et al. 2004), one of the additional questions

asked by this study was, 'What is the natural reliability of grades produced by the graduate teaching assistants? Does training with the Rubric seem to improve grading consistency?'

In general, graduate students under natural conditions produced lower reliability scores than trained raters (Table 10), though natural rater reliability was still comparable to other scores produced by trained raters in the literature (compare to Table 6). Reliability was similar for some criteria (e.g. Introduction: Context and Introduction: Accuracy, Methods: Experimental Design, Primary Literature and the Total Score) but varied noticeably between natural and trained raters for others (e.g. Results: Statistics; Figure 2). Thus, a few hours of training did cause a noticeable increase in the reliability of most scores produced by graduate teaching assistants (Figure 2). Comments by graduate students suggest that this increase in consistency is likely due to the portion of the training session wherein raters scored three example papers and discussed them:

Table 10. Effect of five hours of training on the reliability of scores given by graduate student teaching assistants.

	Trained raters			Natural raters		
Course	one rater	two raters	three raters	one rater	two raters	three raters
101	0.66	0.79	0.85	0.51	0.68	0.76
102	0.66	0.79	0.85	0.57	0.73	0.80
301	0.66	0.79	0.85	0.68	0.81	—[a]

[a]No third rater available (see Section 'Methodology' for more details).
Note: Within a course, trained and natural raters scored the same papers (n = 142 papers total).

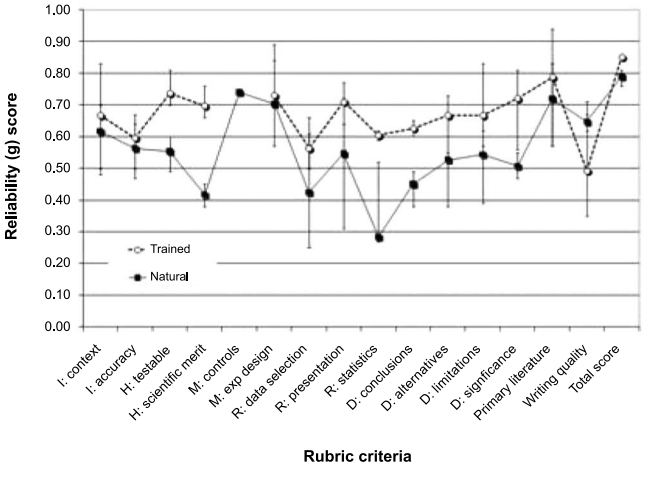

Figure 2. Comparison of the reliability of individual criterion scores for trained versus natural raters.
Note: Data points are the average three-rater reliability across all three courses (n = 142 papers) except for M: Controls which is a single data point from BIOL 101: Natural Raters only (reliability = 0.74, n = 49 papers). The top of each bar indicates the maximum reliability and the lower bar the minimum reliability achieved by that type of rater (bars are not standard error bars).

> The practice lab reports were very beneficial. Until we looked at what you guys scored [on the exemplars] we weren't too sure of what applied for criteria for example.

As previously mentioned, discussion of discrepant scores until raters reached consensus served to calibrate rater's interpretive frameworks. Thus, even though scoring was done independently, the initial discussion and calibration during training likely increased their tendency to interpret and apply the criteria in similar ways. Instructors may therefore wish to incorporate the use of a few exemplar papers and a similar discussion of discrepant scores with their teaching assistants prior to the grading of major course assignments.

No studies were found which directly tested the reliability of graduate student scoring. The only published study found to date that investigated the scores awarded by graduate student instructors in an undergraduate science class compared the relative rankings. Kelly and Takao (2002) compared the relative rankings of research papers in a university oceanography class using the scores assigned by teaching assistants versus those assigned by the instructor (using the rubric developed for course grading) and found agreement only as to which paper was the worst. When the rank orders of the student papers produced by the graduate teaching assistants were compared with those produced by trained raters using an epistemic rubric, there was also little correlation (r = 0.12; Kelly and Takao 2002). The natural reliability of teaching assistants in our study thus appears to be notably higher (g = 0.76 and 0.80 for three raters and 0.81 for two raters).

A likely explanation for this finding is that teaching assistants under natural conditions in this study may actually have received more pedagogical training and support for consistency in grading than did the teaching assistants in Kelly and Takao's study. The graduate students who participated as natural raters in this study had all taught in the introductory biology course at least once at some point in the past. They consequently had at least a year of teaching experience that had included weekly faculty supervision and grading using rubrics. Thus, our natural raters may have greater experience with applying criteria to laboratory reports than did the teaching assistants reported in Kelly and Takao's study. The level of support described here for teaching assistants appears to be greater than that reported for many other institutions, which commonly provide just a single orientation or training session prior to a graduate student's first teaching experience (Boyer Commission 2001). Training in how to teach is not even considered in discussions of the quality of many doctoral programmes (e.g. Carnegie Initiative on the Doctorate 2001; Mervis 2000), and graduate teaching assistants appear to commonly work autonomously with little pedagogical support or training (Davis and Fiske 2001; Golde 2001; Luft et al. 2004). Therefore, other institutions that have not previously incorporated calibrating tools such as a universal rubric may see larger jumps in reliability and consistency from natural conditions to trained conditions if they chose to incorporate this or other rubrics.

Graduate student teaching assistants' perceptions of the Rubric and its effectiveness

Because graduate teaching assistants' perceptions of the utility of a tool are likely to impact their use of the tool, graduate student perceptions of the Rubric were surveyed anonymously immediately after the completion of the Rubric training and scoring sessions. Most raters reported that the concreteness and specificity of the Rubric made scoring easier than grading without a rubric:

> It highlights several categories that are expected in scientific writing and allows for fairly easy and unbiased assessment of whether students are competent in these areas across their academic years.

> Straight forward; very well organised/formatted document – manageable and efficient

They also often felt that training was useful and that it would be beneficial for science departments to provide such training to their teaching assistants (TAs):

> TA orientation should have at least an hour dedicated to working with and calibrating with the rubric – a must if scientific writing is to be a major objective of the department.

> Absolutely [training] should be given to new TAs. Specific instructions will help them grade more consistently – as in how to handle specific errors, specific misconceptions, etc.

If departments choose to provide training to graduate students, the use of a rubric and exemplar papers is therefore recommended as a minimum component of that training. When asked if they would incorporate elements of the Rubric into their own assignments in the future, most graduate students replied affirmatively. Specific comments indicated that they already did use such criteria or listed specific criteria on which they thought the students should focus. Overall students' comments indicated a desire for more incorporation of Rubric elements into departmental courses:

> Believe it or not, this scoring experience really makes me wish I TA'ed a writing intensive course! I would love the opportunity to help my students develop into expert writers over the semester and would definitely use this tool to do so.

Suggestions for improving the Rubric focused mostly on adding additional criteria for various elements the graduate students thought were missing or for giving greater detail in the Rubric about how to handle specific scoring situations. Notably, there were no suggestions to shorten the Rubric.

Cautions and limitations

Instructors and programme evaluators are cautioned to view the Rubric as a tool rather than an answer. Post-hoc application of the Rubric is likely to be unproductive. There must be intentional and conscious alignment between curriculum goals, course design, assignment details and Rubric criteria in order for students to reasonably develop the desired skills over time and for laboratory report or other science writing scores to consequently show meaningful improvement. Without such intentional coordination, the Rubric scores will mostly return information that reflects the misalignments among these factors. Within a course, instructors are specifically encouraged to select Rubric criteria that are directly relevant to their instructional goals *prior* to the development of the assignment. Rubric criteria must be a natural fit for the assignment or the assignment must be designed to address those criteria. Instructional practices must also consistently value and support those criteria (i.e. students need opportunities to practice the desired skills).

Lastly, as demonstrated in Table 7, the reliability of any single criterion can vary, but the reliability of the cumulative score derived from multiple criteria ('Total score' in Table 7) is consistently high. Therefore, instructors are encouraged to incorporate multiple criteria into each assignment and discouraged from using a single criterion in isolation. Use of multiple criteria and summation of scores into a cumulative total provides increased confidence in the reliability of the overall score for the assignment.

Conclusions and recommendations

The Rubric for Science Writing was demonstrated to be a reliable metric in the hands of graduate student teaching assistants and effectively assessed student performance over a variety of biological course contents. On average, students in the reported courses scored within the 'novice' performance level as expected. Future work will focus on capturing a great variety of student writing from upper division courses and mentored undergraduate research experiences. These samples would be expected to score in the 'intermediate' to 'proficient' categories.

The Rubric allows instructors to identify the skills students are acquiring and skills that remain challenging. Such information is necessary if teaching is to be responsive to students' needs. Further, the Rubric illustrated that when criteria are incorporated in multiple courses over time, students' performances will improve.

Application of the Rubric to multiple course assignments also highlighted gaps and misalignments between assignment expectations, desired student performance and curriculum goals. When core expectations were dropped from an assignment, student performance often declined in those areas. This linkage between a consistent scope and sequence for curriculum expectations and attempts to measure student skill development over time cannot be over-emphasised. It is critical that expectations for student performances be consistent over time and that students be given opportunities to learn and practice those skills repeatedly in order for gains to be seen. Scores must be interpreted within the context of course assignments in order to distinguish between areas where students are given opportunities to learn, but truly are not gaining, and instances where the performance of that skill was simply excluded from the assignment. This suggests that the Rubric is also a useful assessment tool for programmatic evaluation as well as being useful in a variety of specific courses. It should be noted that when instructors incorporate rubric criteria into their course assignments, they will naturally adjust the point values assigned to the criteria and expected levels of student performance so that resulting scores translate into appropriate grades. Thus, if the Rubric is used to assess student progress over multiple courses, papers will need to be scored independent of course grading. Explicitly, grading within a course and scoring for programmatic assessment are distinct processes with different purposes; 'A' papers produced by freshman in an introductory course likely will not (and should not) score the same as 'A' papers produced by seniors in an upper division course. We therefore recommend that departments thus collect and electronically store student lab reports from a variety of courses using any number of course software tools. As many institutions now use peer review, plagiarism or course management software packages, collecting portfolios of student work is simply a matter of maintaining archives. Student papers can then be sampled as needed for programmatic assessment.

We recommend that departments support their curriculum goals with relevant performance criteria and that instructors then frequently incorporate those criteria into their assignments. Ideally, instructional goals should span multiple courses and expectations for student performance should be consistently defined by rubric criteria and developed throughout those educational experiences. While these findings do not directly address this issue, we also recommend that instructors share any rubrics openly with students as the descriptions of performance levels can improve student understanding of criteria and expectations. Allowing students to mark each others' papers using the same criteria used by the instructors can also improve student understanding of instructional goals (Timmerman and Strickland 2009).

Graduate student raters produced scores with reliability values comparable or exceeding similar situations in the literature. However, because of the long-standing use of the Rubric in the introductory biology courses at this institution, even the untrained graduate students engaged in this study likely had more experience scoring with the Rubric than would be typical at other institutions. Nonetheless, graduate teaching assistants participating in the training recommended that such training be provided to all incoming graduate students. A key element of the training for improving reliability is likely the use of exemplar papers and discussion to reach consensus (Johnson, Penny, and Gordon 2009). We therefore recommend that instructional staff and teaching assistants score exemplars together and discuss how they will interpret and apply the rubric criteria to student work. We also recommend that in large, multiple section courses where grades are assigned by teams of graduate students, the scoring of exemplars and discussion of criteria occur for each assignment, especially whenever new instructors are present.

Acknowledgements

Many thanks to the numerous faculty who provided feedback on the Rubric and to Sarah A. Woodin in particular for her support of the project. We also extend our gratitude to the dozens of biology graduate students who worked so hard teaching in the laboratories and who participated in both the pilot and the full reliability studies. Finally, our thanks to our colleagues Rebecca Higgins and Michelle Vieyra for their continued input and support of the Rubric at other institutions.

Notes on contributors

Briana Timmerman is associate dean of the Honors College, University of South Carolina and a research associate professor in the Department of Biological Sciences. She enjoys asking how undergraduates and graduate students become effective practitioners of research.

Denise Strickland is a research associate in the Department of Educational Studies with a publication record in ecological research, multi-national teaching experience and an avid interest in how people communicate scientific ideas.

Robert Johnson is a professor of applied measurement at the University of South Carolina. He is author of *Assessing performance: Developing, scoring, and validating performance tasks.*

John Payne is an education associate in the Office of Exceptional Children at the South Carolina Department of Education. He is completing his PhD in Foundations of Education at the University of South Carolina.

References

Aleven, V., and K.R. Koedinger. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science* 26, no. 2: 147–79.

Alters, B.J., and C.E. Nelson. 2002. Perspective: Teaching evolution in higher education. *Evolution* 56, no. 10: 1891–901.

American Association for the Advancement of Science. 1993. *Benchmarks for science literacy.* New York: Oxford University Press.

AERA (American Educational Research Association), APA (American Psychological Association), and NCME (National Council on Measurement in Education). 1999. *Standards for educational and psychological testing.* Washington, DC: AERA Publications.

Baker, E.L., J. Abedi, R.L. Linn, and D. Niemi. 1995. Dimensionality and generalizabilty of domain-independent performance assessments. *Journal of Educational Research* 89, no. 4: 197–205.

Basey, J.M., T.N. Mendelow, and C.N. Ramos. 2000. Current trends of community college lab curricula in biology: An analysis of inquiry, technology, and content. *Journal of Biological Education* 34, no. 2: 80–6.

Baxter, G., R. Shavelson, S. Goldman, and J. Pine. 1992. Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement* 29, no. 1: 1–17.

Black, P., and D. Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5, no. 1: 7–74.

Boyer Commission. 1998. *Reinventing undergraduate education: A blueprint for America's research universities.* Washington, DC: Carnegie Foundation for the Advancement of Teaching.

Boyer Commission. 2001. *Reinventing undergraduate education. Three years after the Boyer Report: Results from a survey of research universities.* Washington, DC: National (Boyer) Commission on Educating Undergraduates in the Research University.

Brennan, R.L. 1992. Generalizability theory. *Educational Measurement: Issue and Practice* 11: 27–34.

Cabrera, A.F., C.L. Colbeck, and P.T. Terenzini. 2001. Developing performance indicators for assessing classroom teaching practices and student learning. *Research in Higher Education* 42, no. 3: 327–52.

Campbell, B., L. Kaunda, S. Allie, A. Buffler, and F. Lubben. 2000. The communication of laboratory investigations by university entrants. *Journal of Research in Science Teaching* 37, no. 8: 839–53.

Carnegie Initiative on the Doctorate. 2001. *Overview of doctoral educational studies and reports: 1990 – present.* Stanford, CA: The Carnegie Foundation for the Advancement of Teaching.

Cho, K., C.D. Schunn, and R.W. Wilson. 2006. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology* 98, no. 4: 891–901.

Cicchetti, D.V. 1991. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences* 14: 119–35.

Clark, D.B., and V.D. Sampson. 2006. *Characteristics of students' argumentation practices when supported by online personally seeded discussions.* Paper presented at the National Association for Research in Science Teaching, April 3–6, in San Francisco, CA.

Committee on Undergraduate Biology Education. 2003. *Bio 2010: Transforming undergraduate education for future research biologists.* Washington, DC: National Academies Press.

Crick, J.E., and R.L. Brennan. 1984. General purpose analysis of variance system (Version 2.2) [Fortran]. Iowa City, IA: American College Testing Program.

Dalkey, N.C. 1969. *The Delphi method: An experimental study of group opinion* (Report RM-5888-PR). Santa Monica, CA: The Rand Corporation.

Davis, G., and P. Fiske. 2001. *The 2000 national doctoral program survey.* Columbia: National Association of Graduate-Professional Students, University of Missouri.

Erduran, S., S. Simon, and J. Osborne. 2004. Tapping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education* 88, no. 6: 915–33.

Feldon, D. 2007. The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review* 19, no. 2: 91–110.

Germann, P.J., and R.J. Aram. 1996. Student performances on the science processes of recording data, analyzing data, drawing conclusions, and providing evidence. *Journal of Research in Science Teaching* 33, no. 7: 773–98.

Golde, C.M. 2001. *Findings of the survey of doctoral education and career preparation: A report to the Preparing Future Faculty program.* Madison, WI: University of Wisconsin.

Haaga, D.A.F. 1993. Peer review of term papers in graduate psychology courses. *Teaching of Psychology* 20, no. 1: 28–32.

Hafner, J., and P. Hafner. 2003. Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education* 25, no. 12: 1509–28.

Halonen, J.S., T. Bosack, S. Clay, M. McCarthy, D.S. Dunn, G.W. Hill, IV, R. McEntarffer, et al. 2003. A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology* 30, no. 3: 196–208.

Hogan, K., B.K. Nastasi, and M. Pressley. 2000. Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition & Instruction* 17, no. 4: 379–432.

Johnson, R.L., J. Penny, S.P. Fisher, and T. Kuhs. 2003. Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education* 16, no. 4: 299–322.

Johnson, R.L., J. Penny, and B. Gordon. 2000. The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education* 13, no. 2: 121–38.

Johnson, R., J. Penny, and B. Gordon. 2009. *Assessing performance: Developing, scoring, and validating performance tasks.* New York: Guilford.

Johnson, R.L., J. Penny, B. Gordon, S.R. Shumate, and S.P. Fisher. 2005. Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly: An International Journal* 2, no. 2: 117–46.

Kelly, G.J., and A. Takao. 2002. Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Science Education* 86, no. 3: 314–42.

Klein, S.P., B.M. Stecher, R.J. Shavelson, D. McCaffrey, T. Ormseth, R.M. Bell, K. Comfort, and A.R. Othman. 1998. Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education* 11, no. 2: 121–38.

Lin, X., and J.D. Lehman. 1999. Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching* 36, no. 7: 837–58.

Linstone, H.A., and M. Turoff, eds. 1975. *The Delphi method: Techniques and applications.* Reading, MA: Addison-Wesley.

Luft, J.A., J.P. Kurdziel, G.H. Roehrig, J. Turner, and J. Wertsch. 2004. Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching* 41, no. 3: 211–33.

Marcoulides, G.A., and M.G. Simkin. 1995. The consistency of peer review in student writing projects. *Journal of Education for Business* 70, no. 4: 220–4.

Marsh, H.W., and S. Ball. 1981. Interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *Journal of Educational Psychology* 73, no. 6: 872–80.

Marsh, H.W., and S. Ball. 1989. The peer review process used to evaluate manuscripts submitted to academic journals: Interjudgmental reliability. *Journal of Experimental Education* 57, no. 2: 151–69.

Marsh, H.W., and P. Bazeley. 1999. Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. *Multivariate Behavioral Research* 34, no. 1: 1–30.

Mervis, J. 2000. Graduate educators struggle to grade themselves. *Science* 287, no. 5453: 568–70.

Moskal, B.M., and J.A. Leydens. 2000. Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation* 7, no. 10, http://PAREonline.net/getvn.asp?v=7&n=10 (accessed September 1, 2009).

National Research Council. 1996. *National science education standards.* Washington, DC: National Academy Press.

National Science Board. 2010. *Science and engineering indicators 2010.* Arlington, VA: National Science Foundation (NSB 10-01).

Novak, J.R., J.L. Herman, and M. Gearhart. 1996. Establishing validity for performance-based assessments: An illustration for collections of student writing. *Journal of Educational Research* 89, no. 4: 220–33.

Penny, J., R.L. Johnson, and B. Gordon. 2000a. The effect of rating augmentation on interrater reliability: An empirical study of a holistic rubric. *Assessing Writing* 7, no. 2: 143–64.

Penny, J., R.L. Johnson, and B. Gordon. 2000b. Using rating augmentation to expand the scale of an analytic rubric. *Journal of Experimental Education* 68, no. 3: 269–87.

Petty, R.E., M.A. Fleming, and L.R. Fabrigar. 1999. The review process at PSPB: Correlates of interreviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin* 25, no. 2: 188–203.

Reeves, G., and L.R. Jauch. 1978. Curriculum development through Delphi. *Research in Higher Education* 8, no. 2: 157–68.

Ruiz-Primo, M., L. Min, C. Ayala, and R. Shavelson. 2004. Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education* 26, no. 12: 1477–506.

Schwarz, C.V., B.J. Reiser, E.A. Davis, L. Kenyon, A. Acher, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik. 2009. Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching* 46, no. 6: 632–54.

Sevian, H., and L. Gonsalves. 2008. Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education* 30, no. 11: 1441–67.

Shannon, D.M., D.J. Twale, and M.S. Moore. 1998. TA teaching effectiveness: The impact of training and teaching experience. *Journal of Higher Education* 69, no. 4: 440–66.

Shavelson, R., and L. Huang. 2003. Responding responsibly to the frenzy to assess learning in higher education. *Change* 35, no. 1: 10–20.

Shavelson, R.J., and N.M. Webb. 1991. *Generalizability theory*, 1st ed. London: Sage.

Tariq, V.N., L.A.J. Stefan, A.C. Butcher, and D.J.A. Heylings. 1998. Developing a new approach to the assessment of project work. *Assessment & Evaluation in Higher Education* 23, no. 3: 221–40.

Timmerman, B., and D.C. Strickland. 2009. Faculty should consider peer review as a means of improving students' scientific reasoning skills. *Journal of the South Carolina Academy of Science* 7, no. 1: 1–7.

Topping, K.J., E.F. Smith, I. Swanson, and A. Elliot. 2000. Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education* 25, no. 2: 149–69.

Volkmann, M.J., and M. Zgagacz. 2004. Learning to teach physics through inquiry: The lived experience of a graduate teaching assistant. *Journal of Research in Science Teaching* 41, no. 6: 584–602.

Willison, J., and K. O'Regan. 2007. Commonly known, commonly not known, totally unknown: A framework for students becoming researchers. *Higher Education Research & Development* 26, no. 4: 393–409.

Yorke, M. 2003. Formative assessment in higher education: Moves toward theory and enhancement of pedagogic practice. *Higher Education* 45, no. 4: 477–501

Appendix 1. Universal rubric for science writing

For more information, please contact: Briana Timmerman Department of Biological Sciences and SC Honors College *timmerman@schc.sc.edu*, 803-777-0822

The rubric was developed as a means of measuring how well students are achieving the stated curriculum goals of the USC Biology Curriculum (http://www.biol.sc.edu/undergrad/curriculum.html) but has been successfully applied in a wide variety of scientific disciplines.

The criteria were selected as the minimum framework one would expect to see in any good scientific communication. The levels of student performance are intended as a roadmap of the probable learning trajectory of a typical undergraduate student over the course of their entire undergraduate career. The 'Proficient' level describes the performance we would hope an exceptional undergraduate or beginning graduate student would achieve when nearing graduation. Faculty mentors and instructors are encouraged to select and use only the criteria and levels of student performance that they feel are relevant to their students and assignment goals. To use the rubric, simply assign point values to each criterion and performance level (e.g. for any given criterion, novice level performance could earn one pt, proficient could earn three pts etc.). A scoring guide (rubric plus examples of student work at each level of performance) has also been developed and is available upon request. Feedback or comments would be most appreciated at the contact information listed above. The Rubric underwent formal reliability testing producing a three rater reliability score of $g = 0.85$ (generalizability analysis) using biology graduate students as raters.

Support for this project was provided by NSF DUE CCLI Award 0410992 and NSF REESE Award 0723686

Criteria	Not addressed	Novice	Intermediate	Proficient
Introduction: Context				
Demonstrates a clear understanding of the 'big picture;' Why is this question important/ interesting in this field? What do we already know? What problem/ question is this research project addressing?	• The importance of the question is not addressed. • How the question relates within the broader context of biology is not addressed.	• The writer provides a generic or vague rationale for the importance of the question. • The writer provides vague or generic references to the broader context of biology.	• The writer provides one explanation of why others would find the topic interesting. • The writer provides some relevant context for the research question(s).	• The writer provides a clear sense of why this knowledge may be of interest to other researchers in that field. • The writer describes the current gaps in our understanding of this field and explains how this research will help fill those gaps.
Introduction: Accuracy and relevance				
Content knowledge is accurate, relevant and provides appropriate background for reader including defining critical terms.	• Background information is missing or contains major inaccuracies. • Background information is accurate, but irrelevant or too disjointed to make relevance clear. • Primary literature references are absent or irrelevant. May contain website or secondary references. **websites or review papers are not primary**	• Background omits information or contains inaccuracies which detract from the major point of the paper. • Background information is overly narrow or overly general (only partially relevant). • Primary literature references, if present, are inadequately explained.	• Background information may contain minor omissions or inaccuracies that do not detract from the major point of the paper. • Background information has the appropriate level of specificity to provide relevant context. • Primary literature references are relevant and adequately explained but few.	• Background information is completely accurate. • Background information has the appropriate level of specificity to provide concise and useful context to aid the reader's understanding. • Primary literature references are relevant, adequately explained, and indicate a reasonable literature search.

Criteria	Not addressed	Novice	Intermediate	Proficient
Hypotheses: Testable and consider alternatives				
Hypotheses are clearly stated, testable and consider plausible alternative explanations.	• No hypothesis is indicated. • The hypothesis is stated but too vague or confused for its value to be determined. • A clearly stated, but not testable hypothesis is provided. • A clearly stated and testable, but trivial hypothesis is provided.	• A single relevant, testable hypothesis is clearly stated. • The hypothesis may be compared with a 'null' alternative which is usually just the absence of the expected result.	• Multiple relevant, testable hypotheses are clearly stated. • Hypotheses address more than one major potential mechanism, explanation or factors for the topic.	• A comprehensive suite of testable hypotheses are clearly stated which, when tested, will distinguish among multiple major factors or potential explanations for the phenomena at hand.
Hypotheses: Scientific merit				
Hypotheses have scientific merit.	• Hypotheses are trivial, obvious, incorrect or completely off topic.	• Hypotheses are plausible and appropriate though likely or clearly taken *directly* from course material.	• Hypotheses indicate a level of understanding beyond the material directly provided to the student in the lab manual or coursework.	• Hypotheses are novel, insightful, or actually have the potential to contribute useful new knowledge to the field.

Criteria	Not addressed	Novice	Intermediate	Proficient
Methods: Controls and replication				
Appropriate controls (including appropriate replication) are present and explained. *If the student designed the experiment:*	• Controls and/or replication are nonexistent. • Controls and/or replication may have been present, but just not described or • Controls and/or replication were described but were inappropriate.	• Controls consider one major relevant factor. • Replication is modest (weak statistical power).	• Controls take most relevant factors into account. • Controls include positive and negative controls if appropriate. • Replication is appropriate (average sample size with reasonable statistical power).	• Controls consider *all* relevant factors. • Controls have become methods of differentiating between multiple hypotheses. • Replication is robust (sample size is larger than average for the type of study).
If the instructor designed the experiment:	• Student fails to mention controls and/or replication or mentions them, but the description or explanation is incomprehensible.	• Student explanations of controls and/or replication are vague, inaccurate or indicate only a rudimentary sense of the need for controls and or replication.	• Student evidences a reasonable sense of why controls/replication matter to this experiment. • Explanations are mostly accurate.	• Explanations of why these controls matter to this experiment are thorough, clear and tied into sections on assumptions and limitations.
Methods: Experimental design				
Experimental design is likely to produce salient and fruitful results (tests the hypotheses posed.) *Methods are:*	• Inappropriate. • Poorly explained/ indecipherable.	• Appropriate. • Clearly explained. • Drawn directly from coursework. • Not modified where appropriate.	• Appropriate. • Clearly explained. • Modified from coursework in appropriate places. • Or drawn directly from a novel source (outside the course).	• Appropriate. • Clearly explained. • A synthesis of multiple previous approaches or an entirely new approach.

Criteria	Not addressed	Novice	Intermediate	Proficient
Results: Data selection				
Data are comprehensive, accurate and relevant.	• Data are too incomplete or haphazard to provide a reasonable basis for testing the hypothesis.	• At least one relevant dataset per hypothesis is provided but some necessary data are missing or inaccurate. • Reader can satisfactorily evaluate some but not all of writer's conclusions.	• Data are relevant, accurate and complete with any gaps being minor. • Reader can fully evaluate whether the hypotheses were supported or rejected with the data provided.	• Data are relevant, accurate and comprehensive. • Reader can fully evaluate validity of writer's conclusions and assumptions. • Data may be synthesized or manipulated in a novel way to provide additional insight.
Results: Data presentation				
Data are summarized in a logical format. Table or graph types are appropriate. Data are properly labeled including units. Graph axes are appropriately labeled and scaled and captions are informative and complete. *Presentation of data:*	• Labels or units are missing which prevent the reader from being able to derive any useful information from the graph. • Presentation of data is in an inappropriate format or graph type. • Captions are confusing or indecipherable.	• Contains some errors in or omissions of labels, scales, units etc., but the reader is able to derive some relevant meaning from each figure. • Is technically correct but inappropriate format prevents the reader from deriving meaning or using it. Captions are missing or inadequate.	• Contains only minor mistakes that do not interfere with the reader's understanding and the figure's meaning is clear without the reader referring to the text. • Graph types or table formats are appropriate for data type. • includes captions that are at least somewhat useful.	• Contains *no* mistakes. • Uses a format or graph type which highlights relationships between the data points or other relevant aspects of the data. • May be elegant, novel, or otherwise allow unusual insight into data. • Has informative, concise and complete captions.

Criteria	Not addressed	Novice	Intermediate	Proficient
Results: Statistical analysis				
Statistical analysis is appropriate for hypotheses tested and appears correctly performed and interpreted with relevant values reported and explained.	• No statistical analysis is performed. • Statistics are provided but are inappropriate, inaccurate or incorrectly performed or interpreted so as to provide no value to the reader.	• Appropriate, accurate descriptive statistics only are provided. • Inferential statistics are provided but either incorrectly performed or interpreted or an inappropriate test was used. • Appropriate, correct inferential statistics are provided, but lack sufficient explanation.	• Appropriate inferential (comparative) statistical analysis is properly performed and reasonably well explained. • Explanation of significant value may be limited or rote (e.g. use of $p<0.05$ only).	• Statistical analysis is appropriate, correct and clearly explained. • Includes a description of what constitutes a significant value and why that value was chosen as the threshold (may choose values beyond $p<0.05$).
Discussion: Conclusions based on data selected				
Conclusion is clearly and logically drawn from data provided. A logical chain of reasoning from hypothesis to data to conclusions is clearly and persuasively explained. Conflicting data, if present, are adequately addressed.	• Conclusions have little or no basis in data provided. • Connections between hypothesis, data and conclusion are non-existent, limited, vague or otherwise insufficient to allow reasonable evaluation of their merit. • Conflicting data are not addressed.	• Conclusions have some direct basis in the data, but may contain some gaps in logic or data or are overly broad. • Connections between hypothesis, data and conclusions are present but weak. • Conflicting or missing data are poorly addressed.	• Conclusions are clearly and logically drawn from and bounded by the data provided with no gaps in logic. • A reasonable and clear chain of logic from hypothesis to data to conclusions is made. • Conclusions attempt to discuss or explain conflicting or missing data.	• Conclusions are completely justified by data. • Connections between hypothesis, data, and conclusions are comprehensive and persuasive. • Conclusions address and logically refute or explain conflicting data. • Synthesis of data in conclusion may generate new insights.

Criteria	Not addressed	Novice	Intermediate	Proficient
Discussion: Alternative explanations				
Alternative explanations are considered and clearly eliminated by data in a persuasive discussion. *Alternative explanations:*	• Are not provided. • Are trivial or irrelevant. • Are mentioned but not discussed or eliminated.	• Are provided in the discussion only. • May include some trivial or irrelevant alternatives. • Discussion addresses some but not all of the alternatives in a reasonable way.	• Some alternative explanations are tested as hypotheses; those not tested are reasonably evaluated in the discussion. • Discussion of alternatives is reasonably complete, uses data where possible and results in at least some alternatives being persuasively dismissed.	• Have become a suite of interrelated hypotheses that are explicitly tested with data. • Discussion and analysis of alternatives is based on data, complete and persuasive with a single clearly supported explanation remaining by the end of the discussion.
Discussion: Limitations of design				
Limitations of the data and/or experimental design and corresponding implications discussed. *Limitations:*	• Are not discussed.	• Are discussed in a trivial way (e.g. 'human error' is the major limitation invoked).	• Are relevant, but not addressed in a comprehensive way. • Conclusions fail to address or overstep the bounds indicated by the limitations.	• Are presented as factors modifying the author's conclusions. • Conclusions take these limitations into account.
Discussion: Significance of research				
Paper gives a clear indication of the significance of the research and its future directions (future research questions).	• Future directions are not addressed. • Significance of the project is not addressed.	• Future directions are vague, implausible (not possible with current technologies or methodologies), trivial or off topic. • Mentions of significance are vague or inappropriate.	• Future directions are useful, but indicate incomplete knowledge of the field (suggest research that has already been done or is improbable with current methodologies). • Significance demonstrates only partial knowledge of field.	• Future directions are salient, plausible and insightful. • Writer clearly explains how this work fills our knowledge gaps and new questions or opportunities that are opened as a result of this work.

Criteria	Not addressed	Novice	Intermediate	Proficient
Use of Primary Literature				
Relevant and reasonably complete discussion of how this research project relates to others' work in the field (scientific context provided). *Primary literature is defined as:* - peer reviewed - reports original data - authors are the people who collected the data. - published by a non-commercial publisher.	• Primary literature references are not included.	• Primary literature references are limited (only one or two primary references in the whole paper). • References to the textbook, lab manual, or websites may occur. • Citations are at least partially correctly formatted. Note that proper format includes a one-to-one correspondence between in-text and end of text references (no references at end that are not in text and vice versa) as well as any citation style currently in use by a relevant biology journal.	• Primary literature references are more extensive (at least one citation for each major concept). • Literature cited is predominantly (≥ 90%) primary literatures. • Primary literature references are used primarily to provide background information and context for conclusions. • Primary literature references.	• Primary literature references indicate an extensive literature search was performed. • Primary literature references frame the question in the introduction by indicating the gaps in current knowledge of the field. • Primary literature references are used in the discussion to make the connections between the writer's work and other research in the field clear. • Primary literature references are properly and accurately cited.

Criteria	Not addressed	Novice	Intermediate	Proficient
Writing quality				
Grammar, word usage and organization facilitate the reader's understanding of the paper. (Some elements inspired by the SC State Dept. Education Extended-Response Scoring Guide for English Language Arts.)	• Grammar and spelling errors detract from the meaning of the paper. • Word usage is frequently confused or incorrect. • Subheadings are not used or poorly used. • Information is presented in a haphazard way.	• Grammar and spelling mistakes do not hinder the meaning of the paper. • General word usage is appropriate, although use of technical language is may have occasional mistakes. • Subheadings are used and aid the reader somewhat. • There is some evidence of an organizational strategy though it may have gaps or repetitions.	• Grammar and spelling have few mistakes. • Word usage is accurate and aids the reader's understanding. • Distinct sections of the paper are delineated by informative subheadings. • A clear organizational strategy is present with a logical progression of ideas.	• Correct grammar and spelling. • Word usage facilitates reader's understanding. • Informative subheadings significantly aid reader's understanding. • A clear organizational strategy is present with a logical progression of ideas. There is evidence of an active planning for presenting information; this paper is easier to read than most.

Acknowledgements

Many thanks to the following people for their time and efforts in developing and/or providing feedback on the rubric in its various stages of development:

USC Department of Biological Sciences Faculty:

Renae Brodie
Susan Carstensen
Erin Connelly
Brian Helmuth
Laurel Hester
Robert Lawther
David Lincoln
Timothy Mousseau
Richard Showman, Curriculum Committee Chair
Dick Vogt
Sally Woodin, Department Chair

USC Biology Graduate Students/Instructors:

Kyle Aveni-Deforge
Sarah Berke
Pamela Brannock
Andre Brock
Lisa Cox
Rebecca Cozart
Debra Davis
Kenny Fernandez
Sierra Jones
Jennifer Jost
Dino Marshalonis
Kenneth Oswald
Suzanne Pickard
Cliff Ramsdell
Sarah Refi
Denise Strickland
Lauren Szathmary
Michelle Vieyra

University of South Carolina:
Claudia Benitez-Nelson, USC Marine Science Program
Robert Johnson, USC College of Education, Department of Educational Psychology, Research, and Foundations
Loren Knapp, Assistant Dean, USC College of Arts and Sciences

Outside USC
Barbara Buckley, Concord Consortium
David Treagust, Curtin University, Australia