European Journal of Applied Sciences – Vol. 11, No. 2 Publication Date: April 25, 2023

DOI:10.14738/aivp.112.14406.

Passonneau, R. J., Koenig, K., Li, Z., & Soddano, J. (2023). The Ideal versus the Real Deal in Assessment of Physics Lab Report Writing . European Journal of Applied Sciences, Vol - 11(2). 626-644.

SERVICES FOR SCIENCE AND EDUCATION

The Ideal versus the Real Deal in Assessment of Physics Lab Report Writing

Rebecca J. Passonneau
Department of Computer Science and Engineering,
Pennsylvania State University, United States

Kathleen Koenig
Department of Physics,
University of Cincinnati, United States

Zhaohui Li
Department of Computer Science and Engineering,
Pennsylvania State University, United States

Josephine Soddano
Department of Computer Science and Engineering,
Pennsylvania State University, United States

ABSTRACT

Effective writing is important for communicating science ideas, and for writing-to-learn in science. This paper investigates lab reports from a large-enrollment college physics course that integrates scientific reasoning and science writing. While analytic rubrics have been shown to define expectations more clearly for students, and to improve reliability of assessment, there has been little investigation of how well analytic rubrics serve students and instructors in large-enrollment science classes. Unsurprisingly, we found that grades administered by teaching assistants (TAs) do not correlate with reliable post-hoc assessments from trained raters. More important, we identified lost learning opportunities for students, and misinformation for instructors about students' progress. We believe our methodology to achieve post-hoc reliability is straightforward enough to be used in classrooms. A key element is the development of finer-grained rubrics for grading that are aligned with the rubrics provided to students to define expectations, but which reduce subjectivity of judgements and grading time. We conclude that the use of dual rubrics, one to elicit independent reasoning from students and one to clarify grading criteria, could improve reliability and accountability of lab report assessment, which could in turn elevate the role of lab reports in the instruction of scientific inquiry.

Keywords: Science writing assessment, Physics lab reports, Analytic rubrics, Writing assessment reliability.

INTRODUCTION

Writing plays a central role in communicating about scientific ideas, experiments and results, yet instructors find it challenging to provide undergraduate science students with rigorous instruction in science writing. This is especially true in the large-enrollment classes that are the norm in bigger public schools. This paper presents a study of a post-hoc reliability assessment of physics lab reports from a large-enrollment college curriculum that integrates several increasingly difficult writing assignments. The curriculum was designed to support the development of scientific reasoning through theory-evidence coordination [1], and was informed by the Science Writing Heuristic (SWH) [2]. A growing body of evidence finds that asking students to put science ideas into writing enhances inquiry-based science instruction (Graham, Kiuhara, and MacKay 2020; Gere et al. 2019; Huerta and Garza 2019; Clabough and Clabough 2016; Timmerman et al. 2011). An important component of learning to write, however, is to provide students with timely, reliable and informative assessments with appropriate feedback [9]–[11]. We investigated the reliability of the original grades assigned to physics lab reports, and time on task to complete the grading. We present an approach that involves the use of an analytic assessment rubric that can improve reliability, timeliness and informativeness of lab report assessment.

An analytic rubric defines the expectations of a writing assignment along multiple dimensions, such as the ability to state a clear hypothesis, to present claims that test the hypothesis, and to give supporting evidence for each claim using experimental results. Each rubric dimension is rated on the same scale. Studies have shown that analytic rubrics can have multiple benefits, including transparency and accountability for students, and reliability of assessment [8], [12], [13]. To achieve reliable grades post-hoc, we developed distinct assessment rubrics with specific criteria for assignment of distinct degrees of partial credit on each rubric dimension. Concurrently, we trained raters until they could apply the assessment rubrics reliably. A comparison of grades assigned by teaching assistants (TAs) and our post-hoc assessments shows the TA grades to be unreliable, with similar time-on-task for both.

We analyzed over 2,000 physics lab reports to address three research questions:

- RQ 1: To what extent do analytic grading rubrics, which are more specific than rubrics provided to students to define lab report expectations, produce reliable assessments?
- RQ 2: How far from reliable were the original grades assigned by TAs?
- RQ 3: What does the reliable assessment reveal about students' science writing?

A critical factor for achieving reliability is that we created distinct assessment rubrics that parallel the original rubrics where expectations for students are defined, but which provided much more detailed and objective criteria for grading. A comparison of the TA and rater effort

appears in the first subsection of our Results section, suggesting that a more specific assessment rubric potentially reduces the time spent on assessment. To address RQ 2, we show concretely how far the TAs' grading behavior is from the reliable post-hoc assessment, presented as the second subsection of our Results section. In our Discussion section, we discuss which rubric dimensions students find most challenging (RQ 3), based on our reliable post-hoc assessment. Reliable assessment supports more meaningful conclusions about trends in student writing, and identification of science ideas students struggle with.

Inconsistency in rubric application is a well-known issue [14] that counterbalances the evidence for the efficacy of rubrics to improve student writing [15]. However, we find little published work on exactly how unreliable classroom grading is, and what the losses might be regarding instructors' ability to adapt classroom practice to the needs of students. Our main objectives are to highlight the potential gains from improved reliability of classroom assessments, along with recommendations for ways to improve reliability of classroom grading.

Science Writing and Assessment

Writing is an important part of science that serves to document and communicate ideas, and in addition, supports science learning [5], [16], [17], and the development of scientific reasoning (SR) skills [18], [19]. Three best practices for incorporating writing into science instruction are (1) the use of analytic rubrics to define student expectations, such as how to construct an argument from evidence [8], [12], [13], (2) frequent opportunities for students to practice writing over extended periods [16], [20], [21], and (3) timely feedback for how well a given piece of writing meets expectations [9]–[11], [22]. We present evidence here for the importance of a fourth criterion, that assessment feedback should also be reliable. In his text on teaching science and engineering [23], Kalman notes that students find it difficult to shift from oral to written discourse. He points out that in conversation, listeners provide feedback that shows a speaker which parts of their discourse are engaging or confusing through explicit comments, or implicit signals such as eye gaze and facial expression. In [9], the authors delineate numerous opportunities for students to receive feedback. They also argue for students and teachers to build *assessment literacy*, such as how to set expectations about the type of feedback students should receive and how they should use it. An important role of a writing rubric is to account to students for each grade point in their assessment, so that students can tackle the next report with a better understanding of how to meet expectations. For a rubric to serve as feedback, however, it must be applied reliably.

Theory-Evidence-Coordination Lab Curriculum

Current education goals include fostering high end skills, such as non-routine problem solving, systems thinking, and critical thinking [24], [25], all of which are foundational for scientific reasoning [26]. Unfortunately, research has shown that students have difficulty applying scientific reasoning (SR) skills to science-related or everyday life contexts [26]–[32]. Informed by research on the development of SR [25], [33], [34], the physics curriculum we investigate here has multiple components. For a series of four increasingly complex investigations to address specific research questions, the components are pre-lab instruction and exercises that target specific SR skills, authentic scaffolded practice of the targeted skills in classroom

experiments conducted by groups of three to four students, and lab report writing to communicate outcomes.

Although multiple research-validated curricula promote learning through conceptual change [35], [36], our labs expand on these and emphasize mathematical modeling while promoting higher order reasoning through the process of theory-evidence-coordination (TEC) (see Figure 1) [1], [37].

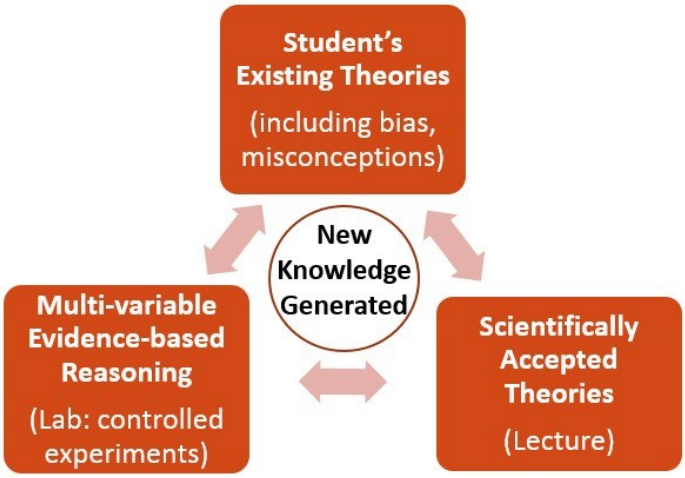


Figure 1. TEC process for knowledge generation.

TEC is an integrative reasoning framework where new knowledge is constructed through the intersection of student's existing theories, data-driven outcomes, and scientifically accepted theories. In the TEC process, students need to master a rich set of SR subskills, including (1) control of variables (COV) reasoning, or the ability to evaluate experimental designs for use of controls, (2) data analytic skills, which extends COV reasoning to identify, manipulate, and evaluate covariation relations from multivariable data; and (3) relating valid evidence (covariation data and relations) with given or hypothesized theoretical claims for explanatory or predictive evaluation under different conditions. Research has shown that students struggle with the TEC process, most likely due to weak SR subskills and limited practice in this type of thinking [26], [38], [39]. Therefore, our labs and lab reports target the SR abilities underlying the TEC process to provide students with rich multi-week investigations. Most of the SR skill development occurs in weekly pre-lab activities in which students are provided with repeated, deliberate practice of select skills within hypothetical scenarios. In the laboratory, students work through the TEC process starting from a research question–such as "What impacts the period of the pendulum?"–followed by brainstorming to elicit students' prior knowledge (theories). Students then generate and test hypotheses with supporting or refuting evidence. This process continues until consistency is reached between the best evaluated hypothesis and supporting evidence. Students are then guided to coordinate their outcomes with scientifically accepted theory, which may lead to new cycles of hypothesis testing, thereby generating new knowledge.

The curriculum has had many successes, especially for retention of underrepresented minorities (URM). Prior to 2013, our labs were cookbook-style with unclear expectations, and

the percentage of students receiving a D, F, or Withdrawing (DFW rate) was 25-28% across all students. Since that time, we have reduced the DFW rates to 6% for non-URM students and 8% for URM students through clearer course expectations and the use of rubrics that clearly define how each assignment is assessed [40], [41].

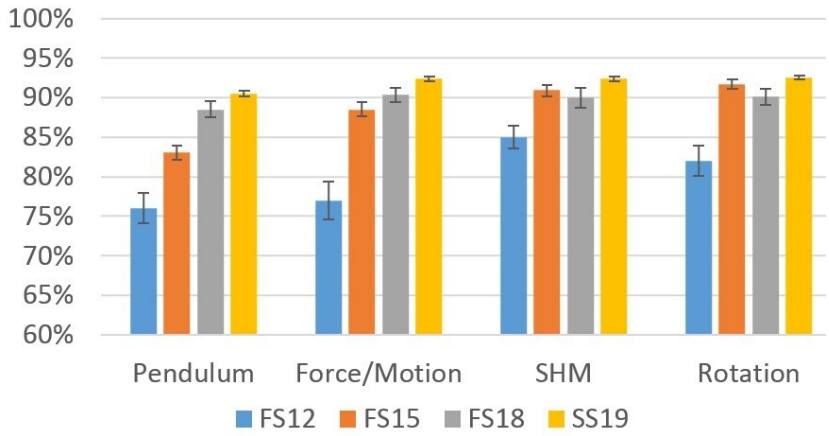


Figure 2. Mean lab report scores across four terms.

However, the effectiveness of the lab report writing has reached a plateau. Figure 3 shows the mean scores of students' lab reports for the four labs over four years. Supporting structures that were added after 2012, such as providing students with writing prompts to guide the TEC process, and rubrics for self-assessment, led to a big initial improvement. The 2019 classroom data suggests that students' lab reports do not continue to improve across the term. However, the apparent trend, as well as our diagnosis, is limited by inconsistency in lab report assessment. More reliable assessment could potentially facilitate improvements in the curriculum, as well as improve student learning.

METHODS

Achieving a reliable post-hoc assessment of the physics lab reports was a necessary step towards measuring the reliability of the TA grades. In this section, we describe the lab reports, the raters, our metrics for measuring reliability, the rubrics, and reliability monitoring during the raters' assessments of the lab reports.

Dataset of Lab Reports

Table 1. Counts of two lab reports by semester and major.

Semester	Life Sci.	Eng.	Subt.
Spr. 2018	-	1	1
Fall 2018	334	-	334
Spr. 2019	164	509	673
Fall 2019	1	71	72
Totals	499	580	1079
Pendulum reports.			

Semester	Life Sci.	Eng.	Subt.
Spr. 2018	-	-	-
Fall 2018	292	-	292
Spr. 2019	164	547	711
Fall 2019	1	1	2
Totals	457	548	1005
Newton's second law reports.			

For this investigation, we used the first two of the four lab reports. The first report addresses whether the period of a pendulum is affected by the length of the string, the mass of the bob on the string, or the angle of release. The second report is about how the acceleration of a system changes when the applied force changes (Newton's second law). Table 1 gives the breakdown of both lab report assignments by semester and by student major. As shown, we re-assessed a total of 1,079 pendulum reports and 1,005 reports on Newton's second law, primarily from the Fall 2018 and Spring 2019 semesters, with a few reports from other semesters. Students majoring in the health or life sciences take the course in their third or fourth year, and their mathematics training typically stops with algebra. These students have often had multiple lab report writing experiences in their biology and chemistry courses. Students majoring in engineering programs take the course in their first or second year, have had at least one semester of calculus, and tend to take the course in the spring.

Rater Recruitment and Training

Four advanced undergraduate majors in computer science with one to three years of prior course work in physics were recruited as raters. Each training phase involved application of an assessment rubric to sets of 5-6 reserved reports, measures of inter-rater agreement, and two group meetings per week to discuss discrepancies and to revise the rubric for more consistent assessment. During training on each report, a bank of consensus examples was created to serve as benchmarks for self-calibration [42]. Training included multiple discussions of examples from actual reports to develop among the raters a common understanding of how to apply the rubric, as recommended in [43].

Metrics

To assess inter-rater agreement among pairs of raters, we used the Pearson correlation coefficient to measure the degree to which two raters applied the same scores per rubric dimension to the same report. As discussed below, we also applied a variant of a kappa score. Pearson correlation measures whether two ordered series of numbers define lines with the same slope. It therefore emphasizes the correlation between pairs of raters across the several dimensions of a rubric. High correlation supports more consistent feedback to students and instructors regarding identification of students' strengths or weaknesses. The critical value of the Pearson correlation, which is dependent on the degrees of freedom (the number of rubric dimensions plus two), indicates whether the achieved value is statistically significant, meaning whether we can reject the null hypothesis that there is no linear correlation between pairs of raters. We calculated the Pearson correlation coefficient ρ for the rubric dimensions on a given lab report, the average ρ over the six pairs of raters for a given report, and the overall average for a batch of student reports that had been assigned at the same time. We refer to this final average of the average correlations for all pairs of raters as the macro-average ρ. A training period ended after meeting three criteria: raters achieved a macro-average $\rho \geq 0.60$, revisions to the rubric were no longer needed, and raters no longer had consequential questions about how to interpret the rubric. The Pearson correlation does not measure closeness of the absolute values assigned by two raters, in contrast to the family of inter-rater agreement metrics that factor out chance agreement, including Cohen's kappa [44], quadratic weighted kappa [45], and Krippendorff's alpha [46]. The kappa-like metrics, however, share the limitation that they are

more open to interpretation, and provide no measure of statistical significance. To get another perspective on inter-rater agreement, we used Krippendorff's alpha, which is a flexible metric that can handle any number of raters, a wide range of data scales, and missing values. We used it with an interval scale for ordered values, which treats the difference between a rating of 5 and 3 as twice the difference between a rating of 5 and 4. One final issue with these chanceadjusted metrics is that on values that raters assign more often, the probability that two raters will agree will be higher, and therefore chance-adjusted agreement will be lower, giving rise to the so-called paradox of kappa [47], [48].

Development of Assessment Rubrics

Two of the authors supervised the modifications to the original rubrics to produce assessment rubrics, and trained the raters. Given that rubric scales ranging from four to six points tend to be the most achievable [49], and that more than seven could lead to cognitive difficulty [50], we chose to convert the original 3-point scale used by the TAs to a 6-point scale (0 to 5). This also accords with best practice for fine-grained analytic rubrics [51], [52]. The size of a rubric scale affects the ability to make meaningful distinctions [49], [52]. Along with widening the scale, we developed specifications for each point on the scale, for each rubric dimension. The assessment rubric for the pendulum report, which has seven dimensions, appears in Figure 3. The two supervising authors developed a first draft of this rubric. In phase one of the pendulum rubric training, each rater working independently assessed the same three reports, and then we discussed divergent ratings. These discussions led to a revision of three rubric dimensions (2 through 4) that pertain to the three experiments testing the effects of mass, length and angle of release on the period of a pendulum. We added the criterion that students should discuss what conclusions can be drawn from the graphs and error bars. A similar process led to a final revision, where we broke both dimensions 5 and 6 down into five distinct elements (see Figure 3). On a final training batch of forty reports, raters reached a new macro-average $\rho = 0.61$. The rubric for the second report on Newton's second law, shown in Figure 4, has one more dimension than the pendulum rubric, and several of the dimensions differ. In the pendulum report, students were to describe three experiments, each investigating how mass, length, or angle of release affected periodicity (dimensions 2-4), whereas for the second report, there is only one experiment (dimension

Inadequate (1)	Inadequate (2)	Needs improvement (3)	Needs improvement (4)	Complete (5)
1. Is able to state the research question for reader clarity.				
Research question is *included* but *incorrectly* stated. *Does not* give an explicit statement of the three variables.	Research question is *included* but *incorrectly* stated. Gives an explicit statement of the three variables.	Research question is *included* and *correctly* stated. Gives an explicit statement of the three variables.	Research question is *included* and *correctly* stated. Gives an explicit but incomplete statement of the three variables.	Research question is *included* and *correctly* stated: **"What affects the period of a pendulum**?" Includes an explicit statement of the three variables: mass, angle of release, and string length.
2. Is able to describe how the experiment with mass as the independent variable addressed the research question and what claim can be made. The mass vs. period graph and error bars are referred to in the discussion.				
Claim about mass is *wrong* (correlates mass to period). Does *not* mention holding other variables constant. Does *not* refer to the plots with error bars. Note: Error bar discussion could also have a correct explanation of standard deviation of uncertainty, one standard error, a particular confidence interval or one other way to explain error bars.	Claim about mass is *wrong* (correlates mass to period). Mentions holding other variables constant. Does *not* refer to the plots with error bars. OR Correct claim about mass (no correlation of mass to period). Does *not* mention holding other variables constant. Refers to the plots with error bars.	Claim about mass is *correct* (no correlation of mass to period). Mentions holding other variables constant. Does *not* refer to the plots with error bars. OR Claim about mass is *wrong* (correlates mass to period). Mentions holding other variables constant. Refers to the plot with error bars.	Claim about mass is *correct* (no correlation of mass to period). Does not mention holding other variables constant. Refers to plot with error bars.	A discussion is included with adequate *reasoning or justification* for *how* the evidence (graphed data) supports the claim (no correlation of mass to period). Mentions holding the other variables constant. Error bars, their size, and the equivalency criterion are discussed with respect to how the mass affects the period.
3. Is able to describe how the experiment with length as the independent variable addressed the research question and what claim can be made. The mass vs. period graph and error bars are referred to in the discussion.				
Claim about length is *wrong* (no correlation of length to period). Does *not* mention holding other variables constant. Does *not* refer to the plots with error bars. Note: Error bar discussion could also have a correct explanation of standard deviation of uncertainty, one standard error, a particular confidence interval or one other way to explain error bars.	Claim about length is *wrong* (no correlation of length to period). Mentions holding other variables constant. Does *not* refer to the plots with error bars. OR Claim about length is *correct* (correlates length to period). Does *not* mention holding other variables constant. Refers to the plots with error bars.	Claim about length is *correct* (correlates length to period). Mentions holding other variables constant or the specific relationship to length. Does *not* refer to the plots with error bars. OR Claim about length is *wrong* (no correlation of length to period). Mentions holding other variables constant. Refers to the plots with error bars.	Claim about length is *correct* (correlates length to period). Refers to plot with error bars. Does *not* mention holding other variables constant or the specific relationship to length.	A discussion is included with adequate *reasoning or justification* for *how* the evidence (graphed data) supports the claim (correlates with length, linear with square root of length or power law relationship). Mentions holding other variables constant. Error bars, their size, and the equivalency criterion are discussed with respect to how the length affects the period.
4. Is able to describe how the experiment with angle of release as the independent variable addressed the research question and what claim can be made. The mass vs. period graph and error bars are referred to in the discussion.				
Claim about angle of release is *wrong* (correlates angle of release to period). Does *not* mention holding other variables constant. Does *not* refer to the plots with error bars. Note: Error bar discussion could also have a correct explanation of standard deviation of uncertainty, one standard error, a particular confidence interval or one other way to explain error bars.	Claim about angle of release is *wrong* (correlates angle of release to period). Mentions holding other variables constant. Does *not* refer to the plots with error bars. OR Claim about angle of release is *correct* (no correlation of angle of release to period). Does *not* mention holding other variables constant. Refers to the plots with error bars.	Claim about angle of release is *correct* (no correlation of angle of release to period). Mentions holding other variables constant. Does *not* refer to the plots with error bars. OR Claim about angle of release is *wrong* (correlates angle of release to period). Mentions holding other variables constant. Refers to the plots with error bars.	Claim about angle of release is *correct* (no correlation of angle of release to period). Does *not* mention holding other variables constant. Refers to the plots with error bars.	A discussion is included with adequate *reasoning or justification* for *how* the evidence (graphed data) supports the claim (no correlation of angle of release to period). Mentions holding other variables constant. Error bars, their size, and the equivalency criterion are discussed with respect to how the angle of release affects the period.
5. Is able to provide the correct theoretical equation for the period of a pendulum and discuss how the mathematical model produced from the lab data supports, or does not support, the theoretical model.				
1) Use correct reference diagram to fit the mathematical model, give correct formula (t T = 2 √L). (1 point) 2) Provide explanation with respect to size of the constant and the exponent. (1 point) 3) Gives the correct theoretical equation (T = 2 π sqr{L/g}). It could also be things like this: T = 2.0061(s/√m)$L^{(0.5)}$. (1 point) 4) Results from curve fitting are included in the discussion; for example, the computing the R-value is an attempt of curve fitting. (1 point) 5) Discuss how the mathematical model produced in lab supports the theoretical model is complete and accurate. (1 point)				
6. Is able to identify random errors and how they were reduced or could be reduced. (Systematic errors are included when applicable.)				
1) Discusses at least 1 random error. (1 point) 2) Discusses at least 1 way to reduce 1. (1 point) 3) Discusses at least 1 systematic error. (1 point) 4) Discusses at least 1 way to reduce systematic error. (1 point) 5) Includes one or more additional random or systematic errors. (1 point)				
7. Is able to identify *constraints* within the experiment and discuss how these may affect the generalizability of the results.				
Mentions incorrect constraints only.	Does not mention one of the constraints of length, mass, and time. Possibly mentions other constraints, like time limitations or poor equipment for the experiment.	Mentions one of the constraints of length, mass, and time. Possibly mentions other constraints, like time limitations or poor equipment for the experiment.	Mentions exactly two of the constraints of length, mass, and time. Possibly mentions other constraints, like time limitations or poor equipment for the experiment.	Mentions all three constraints: length is measured only up to a certain amount; mass is measured only up to a certain amount; time limitations on experiments. These can lead to not enough granularity in mass or length measurements.

Figure 4. Pendulum report assessment rubric: Seven dimensions.

Dimensions 3 and 4 instead asked students to discuss the findings of other student teams. Similar to the pendulum rubric, dimension 5 asked students for the theoretical equation accounting for the effect of force on acceleration. Dimension 6 asked students to provide a second theoretical equation, to account for the way multiple forces would affect the acceleration of a system.

Inadequate (1)	Inadequate (2)	Needs improvement (3)	Needs improvement (4)	Complete (5)
1. Is able to restate the research question for reader clarity.				
Includes an *incorrect* or *imprecise* statement of the research question.	Includes a *correct* but overly *general* statement of the research question that does *not* relate to the design of the experiment (i.e., what variables are manipulated).	Includes a specific but *brief* statement of the research question that explicitly mentions investigating how the application of force changes acceleration, but *without* going into *detail* about how the experiment is designed to answer this question.	Includes a *specific* statement of the research question that explicitly mentions investigating how the application of force changes acceleration, with a *general* explanation of how the experiment is designed to answer this question	Research question is *included* and *correctly* stated: **"How does the acceleration of a system change when the applied force changes?"** Includes a detailed explanation of how the experiment is designed to answer this question that indicates what variables are manipulated, e.g., by stating explicitly what the dependent and independent variables are
2. Is able to describe how the experiment with **hanging weight** as the independent variable addressed the research question and what claim can be made. The acceleration vs. hanging weight graph and its error bars are referred to in the discussion.				
1) Mentions the mass of the system is held constant. (1pt) 2) States the claim about how the hanging weight affects the acceleration: **as the applied force increases, the acceleration increases**. (1pt) 3) Refers to the plots with error bars. (1pt) 4) Gives adequate reasoning or justification for how the evidence (graphed data) supports the claim. (2pt) # Error bars could also be correct explanation of standard deviation of uncertainty, one standard error, a particular confidence interval or things that can explain error bar. Large error bars indicate a lack of confidence in the actual measurements)				
3. Is able to describe how the findings of a group different from the author's group either support or refute the author's group's results and conclusions.				
The other group's findings are clearly presented. Discrepancies and agreements between the groups' data and claims are stated and discussed. 1) Mentions another group's claim or hypothesis. (1pt) 2) Mentions another group's formula. (1pt) 3) Mentions another group's data (unit or other stuff). (1pt) 4) Discusses the discrepancies and agreements between the groups' data and claims. (2pts, or 1pt for a few details)				
4. Is able to describe how the findings of a second group different from the author's group either support or refute the author's group's results and conclusions.				
The other group's findings are clearly presented. Discrepancies and agreements between the groups' data and claims are stated and discussed. 1) Mentions another group's claim or hypothesis. (1pt) 2) Mentions another group's formula. (1pt) 3) Mentions another group's data (unit or other stuff). (1pt) 4) Discusses the discrepancies and agreements between the groups' data and claims. (2pts, or 1pt for a few details)				
5. Is able to provide the correct theoretical equation for the acceleration of a system by a single force and discuss how the experimental mathematical model supports, or does not support, the theoretical model.				
1) Theoretical equation $a=F/m_{sys}$ is provided. (1pt) 2) Mathematical model a=C1*w+C2 is provided. (1pt) 3) The discussion for how the experimental mathematical model supports the theoretical model is complete and accurate. (1pt) 4) Includes results from curve fitting. (1pt) 5) How the system mass relates to the fitting parameters are included in the discussion. (1pt) # How confident are you in the mathematical model provided by Excel? Comment on how well the trendline passes through the set of plotted data points. Cite the R2 value as well and discuss what this value indicates to you.				
6. Is able to provide the correct enhanced theoretical equation for the acceleration of a system by more than one force and discuss how the experimental mathematical model supports, or does not support, the theoretical model.				
1) Enhanced theoretical equation $a=F/m_{sys} + F_{ext}/m_{sys}$ is provided. (1pt) 2) Enhanced Mathematical model a=C1*w+C3*w is provided. (1pt) 3) The discussion for how the experimental mathematical model supports the theoretical model is complete and accurate. (1pt) 4) How the extra force(s) relate(s) to the fitting parameters are included in the discussion. (2pts; 1pt for pointing out the force, 1pt for stating the result of that external force). # For those factor(s) that have impact, describe how each affects the numerical values in your mathematical model. # For instance, discuss how the tilt of the air track affects the sign and magnitude of C3. Also discuss how the retarding force affects the magnitude of C3.				
7. Is able to identify random errors and how they were reduced or could be reduced. (Systematic errors are included when applicable.)				
1) Discusses one random error. (1 pt) 2) Discusses one way to reduce random error. (1 pt) 3) Discusses one systematic error. (1 pt) 4) Discusses one way to reduce systematic error. (1 pt) 5) Includes one or more additional random or systematic errors. (1 pt)				
8. Is able to identify constraints within the experiment and discuss how these may affect the generalizability of the results.				
Identifies incorrect constraints.	An attempt is made to identify constraints, but the discussion is missing for how these may affect the generalizability of the results.	Gives some valid constraints, and explains how they may affect the generalizability of the results.	At least one major constraint is mentioned, and an explanation is given about how it limits generalizability.	At least two of three major constraints are identified and included in a discussion for how they may affect the generalizability of the results: 1) Length of track or limits the measurements that can be taken for cart acceleration. 2) Maximum mass of cart or 3) force applied are based on materials provided.

Figure 5. Newton's second law assessment rubric: Eight dimensions.

Rater Reliability

To perform the post-hoc assessment, after raters had been trained, we assigned separate batches to each rater to assess independently. Unknown to raters, each batch had a small

random subset that was assessed by all four raters for monitoring their reliability throughout the assessment. We continued meeting regularly with the raters to discuss any patterns of disagreement. Table 2 for the pendulum report shows the size of each batch in the Count column (raters were given latitude in choosing their batch size), followed by the number of reports that all four raters assessed in common (Reli.), the average time per report (in tenths of hours, Hr./Rep.), and the macro-averaged ρ for the reports assessed in common. By batch 2, the raters reached macro-average ρ above 0.70, which they maintained through all remaining batches. The correlations thus exceeded the critical value of 0.67 for statistical significance at the 0.05% level (df=5).[1]

Table 2. Reliability monitoring while assessing the pendulum reports.

Batch	Count	Reli. Subset	Hr./Rep.	Mac. Avg. ρ	Avg. α-interv.
1	50	10	0.14	0.62	0.51
2	80	10	0.14	0.73	0.65
3	70	15	0.11	0.76	0.65
4	80	15	0.11	0.77	0.65
5	275	15	0.08	0.72	0.59

The last column of Table 2 reports average Krippendorff's alpha (interval scale) for each batch of the pendulum reports. Interpretation of kappa-like scores is subjective. Scores between 0.41 and 0.60 are considered moderate in agreement, with scores above that having substantial agreement, according to [53]. As shown here, the raters had high average alpha scores on batches 2-4 and moderate alpha on batches 1 and 5. Recall that alpha values for rating values that occur frequently will be lower, due to the paradox of kappa. We observed that the data was skewed towards scores of "5" on dimension 1 (68.0% of the ratings) and dimension 6 (41.0% of the ratings).

Table 3. Reliability monitoring while assessing the Newton's second law reports.

Batch	Count	Reli. Subset	Hr./Rep.	Mac. Avg. ρ	Avg. α-interv.
1	40	10	0.11	0.57	0.44
2	45	10	0.13	0.78	0.69
3	40	10	0.11	0.50	0.56
4	45	10	0.12	0.72	0.60
5	50	10	0.11	0.76	0.67
6	50	10	0.11	0.72	0.62
7	50	21	0.11	0.75	0.63

[1] The total number of reports assessed in Table 2 is $\sum_i(4 \times (|\text{Counti}| - |\text{Relii}|) + |\text{Relii}|))$ for $i \in [1:5] = 1085$. For six of these reports, we could not recover the TA grades, so these are not included in Table 1. [2] One of the raters was unavailable for batch 3.

For the second set of lab reports, we followed the same procedure to revise the rubric and train raters. For this assessment rubric, fewer revisions were needed due to our greater experience. Table 3 shows that during the assessment phase, raters consistently maintained $\rho \geq 0.65$, apart from batch 3.2.[2] The critical value for statistical significance at the 0.05% level for the 8dimension rubric is $\rho \geq 0.63$. The Krippendorff's α values are high for batches 2 and 4-7, despite

highly skewed data where 68% of the time, raters assigned scores of '0' for rubric dimension 6.[2]

Ethics Approval and Consent

The research reported in this manuscript is a retrospective study, and ethical approval was sought prior to starting the study. The ethics committee that reviewed the proposed research was the Institutional Review Board of the University of Cincinnati (FWA #000003152). The IRB determined that the proposed activity was not research involving human subjects. The data involved was historical and de-identified. The IRB ID provided during the review was 20200172.

RESULTS Comparison of TA and Rater Effort

The most costly step in the reliability study was the iterative process of rubric revision and rater training. On average, each rater spent 5.67 hours on assessment of reports during training. Raters also spent about 4 hours in joint review meetings among raters and researchers to discuss divergent scores assigned by raters, and to revise the rubric. Once the raters were fully trained, the average time for trained raters to assess a single report was approximately six and a half minutes, compared with an estimated ten minutes per report that TAs spent. We speculate that TAs could achieve greater reliability during classroom grading with a few straightforward changes. First, TAs could use the more objective assessment rubrics. They could also be given a training session for each rubric, along with calibration examples. The extra hours of TA training would likely be balanced out by greater efficiency per report. In addition, training in the use of the rubrics could give the TAs greater confidence that their efforts could have a positive impact on students, and thus increase their satisfaction.

Reliability of TA Grading

To make the TA grades and raters' scores commensurate, we converted all scores to percentages. The TAs graded lab reports that included sections describing experiments and results along with data tables and plots that had all been previously assessed, and included in the final reports for cross reference. Given our focus on students' scientific writing skills, our raters assessed only the discussion and conclusion sections. The TA mean percentages of the relevant sections on both reports are much higher than those of the same sections scored by the raters, as shown in Table 4. P-values of t-tests comparing the two pairs of means are effectively p=0.00. The low

[2] The total number of reports assessed in Table 3 is $\sum_i(4 \times (|\text{Counti}| - |\text{Relii}|)) + |\text{Relii}|))$ for $i \in [1:7] = 1057$. We could recover the TA grades for 1005 of these; see Table 1.

Pearson correlations of TA grades and rater assessments for the pendulum reports (ρ=0.32) and the Newton's second law reports (ρ=0.19) show that the TA grades do not correlate with the rater assessments.

Table 4. Comparison of TA and rater assessments.

TA Mean (sd)	Rater Mean (sd)
0.89 (0.11)	0.59 (0.17)
T-test	$\rho = 0.32$ p = 0.0000

(a) Pendulum reports.

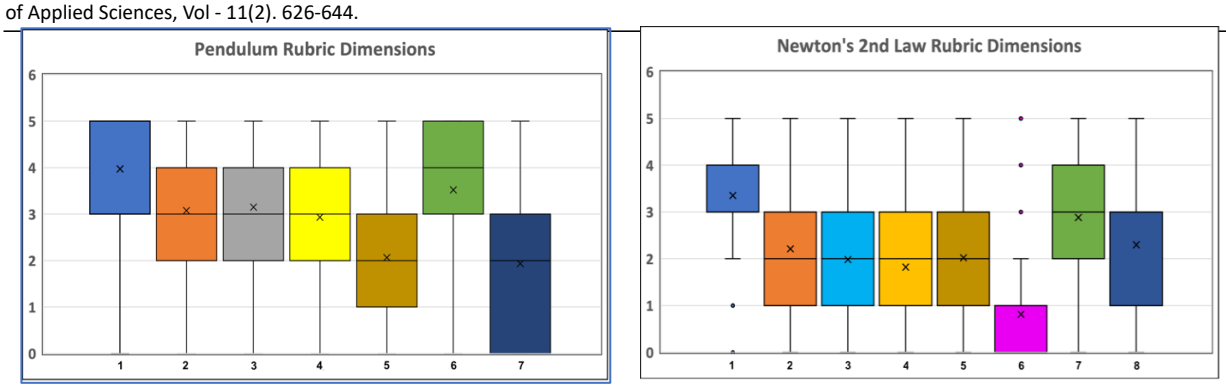
TA Mean (sd)	Rater Mean (sd)
0.92 (0.09)	0.43 (0.16)
T-test	$\rho = 0.19$ p = 0.0000

(b) Newton's second law reports.

Comparison of the TA grades versus the rater assessments provides a picture of the impact on instruction. Although it was a more challenging report, the average TA grade assigned to the Newton's second law report was 3% higher than the grade average TAs gave the first report. In contrast, the average rater score on the second more challenging report was 30% lower than on the first report. Section 7 includes measures of average student performance on different rubric dimensions for both reports, which reveals which aspects of the reports are most difficult for students.

DISCUSSION

Reliable assessment is a necessary precondition to identify the strengths and weaknesses in student writing. It could benefit student learning if students engage with the feedback about how well they performed on each dimension of a lab report. It could also inform how instructors plan subsequent class meetings. Finally, as observed in the introduction, it could inform revisions to a curriculum. Here we point to observations that emerge about students' strengths and weaknesses. We also report an apparent TA bias that favors the engineering majors.



(a) Pendulum report. (b) Newton's second law report.

Figure 6. Breakdown of student performance by rubric dimension.

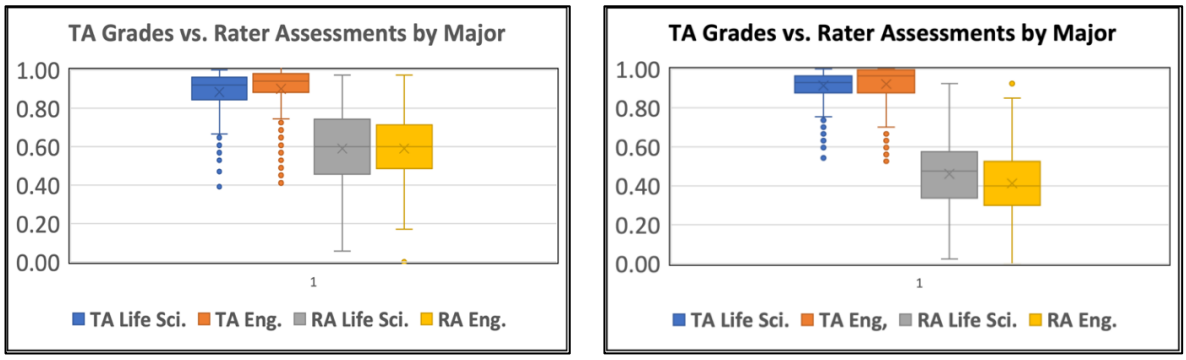
Figure 6 shows box and whisker plots of the mean points achieved on each rubric dimension on both reports on the reliable assessments. Sameness of color coding across the two plots reflects rubric dimensions that appear in both reports. Overall, students did better on the pendulum report. We see a similar trend for the dimensions that recur, thus students did best on stating the research question (dimension 1 in both reports), and explaining how random versus systematic errors affect the results

Pendulum Report, Dimension 3: Experiment about length of string	
Score	Student text excerpts
5	The third experiment investigated whether or not the length of the string affects the period of the pendulum. The graph of length of string vs. period of the pendulum shows points that slope upward, proving there is a positive correlation between the length of the string and the period of the pendulum. This claim holds true when the angle of release of the pendulum is held constant or under 20 degrees. **The error bars on the graph do not overlap or touch, so the trend is not due to error.**
3	To test the string length factor, the angle of release and weight on string were kept constant while the length of string was varied across all the trials. The claim was made that the string length was directly relate (sic) to the period because increased length led to increased period.
Newton Report, Dimension 3: Findings of a Different Group	
Score	Student text excerpts
3.75	In Group 3, they had made the same hypothesis we did, that the applied force will affect the acceleration of the system. . . . they had a experimental model equation of y=5.0587x-0.0419. This equation and their data supports our claim, of the applied force affecting the acceleration. . . . All our experimental model equations have the same positive linear relationship between the applied force and the acceleration. **Our data is different from each of the other groups because all the groups had a different system mass. Thus, giving us all different accelerations and experimental model equations.**
2.25	As one can see our different values for the three trials of the experiment are similar to those of the other groups, validating our data, the differences van be attributed to random error primarily, because everyone did only one iteration of each run, meaning that no averages were taken . . .

Figure 7. Excerpts from pairs of high and medium scoring lab reports.

(dimension 6 in the pendulum report, dimension 7 in the Newton's second law report). Students did better in the first report on describing the three pendulum experiments (dimensions 2-4) than describing the Newton's second law experiment (dimension 2). In the second report, students did similarly well at comparing their results to those from other teams (dimensions 3 and 4) as they did on describing the experiment (dimension 2). They did only moderately well in both reports on providing the correct theoretical equation (dimension 5, both reports). On

the second report, they did poorly on providing an enhanced theoretical equation that deals with multiple forces rather than a single force (dimension 6). Students showed improvement in explaining experimental constraints that might affect the generality of results. On the pendulum report, this is dimension 7, where the average was $\mu = 1.94$ ($\sigma = 1.43$); in the second report, this is dimension 8, where the average score was $\mu = 2.30$ ($\sigma = 1.58$). Figure 6 illustrates contrasting pairs of excerpts from both reports, alongside the average of the four raters' scores. The top half shows a high versus low scoring passage that addresses dimension 3 of the pendulum rubric. In the first passage, mass vs. period graph and error bars are referred to in the discussion (boldface font). The second states a correct claim but does not refer to the error bars. The bottom half of Figure 6 shows a pair of passages that addresses dimension 3 of the second rubric: *Is able to describe how the findings of a group different from the author's group either support or refute the author's group's results and conclusions*. The higher scoring excerpt goes into detail about the reasons for differences in the empirical equations, while the lower scoring excerpt provides no details and little reasoning.



(a) **Pendulum report.** ***(b)*** **Newton's second law report.**

Figure 8. Whisker plot comparison of TA grades and rater assessments by major.

Recall from section 4 that students represent two majors, different seniority in school, different mathematics training, and different prior experience with lab report writing. Figure 7 compares the two pools of students, based on the TA grades versus the rater assessment. According to the TA grades, engineering majors perform slightly better than the life science majors on the pendulum report (ANOVA yields $p=0.0120$), but had no significant difference in mean scores on the Newton's 2nd law report. In contrast the raters' scores show no significant difference by major on the pendulum report, whereas on the more challenging second report, the relative performance by major is flipped: the health or life sciences majors' have a mean score of 0.46 compared with that of the engineering majors of 0.41, which is significantly better (ANOVA yields $p=2.6 \times 10-6$). Thus, the TA grades reveal a bias that favors the engineering students. Our comparison of TA grades to a post-hoc reliable assessment of lab reports points to lost opportunities for student learning, for instructors' understanding of student achievement, and for curriculum revision, due to rubric-based assessments that are not reliable. Our investigation of the resources needed to achieve reliability suggest that highly specific assessment rubrics could simultaneously reduce time to provide feedback, improve assessment reliability, and enhance accountability of students and TAs alike. It might also reduce bias.

We speculate that the use of distinct assessment rubrics could also improve students' understanding of how they met or failed to meet expectations. There has been little investigation of the role of feedback to students on science writing, including the question of whether rubric-based grades are an important component of feedback. In [9], it is argued that there are many opportunities for students to get feedback during a course, and that these should be co-ordinated with one another. They further argue for improved assessment literacy (see above) on the part of instructors and students alike. For this to occur, we believe there is a need for more research on different forms of feedback, and how students engage with feedback. In our future work, we hope to investigate students' engagement with feedback in the form of the kind of assessment rubric discussed here. Although the rubrics presented here are for a specific curriculum, we believe our results could benefit science writing in general as they provide evidence for the use of dual analytic rubrics. Here, one rubric would define the expectations for the students in a general way, and a second one, aligned with the first, would be used for assessment through detailed and objective specifications at each point increment for each analytic dimension. Providing students with an assessment combined with the grading rubric would also serve as detailed feedback regarding ways to improve their lab report writing. We speculate that reliability, consistency, and efficiency of TA grades in STEM curricula could be improved through utilization of rubrics specifically designed for assessment and feedback. We hope to investigate the use of dual rubrics in classroom instruction in our future work.

CONCLUSION

No one recipe for integration of writing into science instruction could possibly apply across the diversity of students, instruction methods, and science disciplines in present day colleges and universities. However, decades of investigation into the science writing heuristic have shown that different kinds of writing exercises for students, from informal reflective writing [23] to highly structured reports [54], each have benefits. For example, a recent meta-analysis of SWH [7] found that most of the significant effects of writing on learning come from studies within a given genre, such as science writing, rather than across genres. Our study highlights the challenges inherent in the assessment of formal writing assignments, a task that college science instructors find difficult [55], and our results suggest a strategy that may better support both instructors (or TAs) and students. We have demonstrated that for lab reports from two semesters of a large-enrollment, introductory physics lab curriculum, the grades assigned by TAs were not reliable, while a reliable post-hoc assessment reveals valuable information about students' strengths and weaknesses. The role of science writing in this lab curriculum is to instruct students in science reasoning in a manner that shows up in their writing, so as to promote further learning. In conclusion, reliable assessment of writing could elevate the role of lab reports in the overall instruction of scientific inquiry.

References

[1] D. Kuhn, "What is Scientific Thinking and How Does It Develop?" in *Blackwell Handbook of Childhood Cognitive Development*, U. Goswami, Ed., Malden, MA, USA: Blackwell Publishers Ltd, 2002, pp. 371–393. doi: 10.1002/9780470996652.ch17.

[2] C. W. Keys, "Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science," *Science Education*, vol. 83, no. 2, pp. 115–130, 1999, doi: 10.1002/(SICI)1098237X(199903)83:2<115:AID-SCE2>3.0.CO;2-Q.

[3] S. Graham, S. A. Kiuhara, and M. MacKay, "The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis," *Review of Educational Research*, vol. 90, no. 2, pp. 179–226, Apr. 2020, doi: 10.3102/0034654320914744.

[4] A. R. Gere, N. Limlamai, E. Wilson, K. MacDougall Saylor, and R. Pugh, "Writing and Conceptual Learning in Science: An Analysis of Assignments," *Written Communication*, vol. 36, no. 1, pp. 99–135, Jan. 2019, doi: 10.1177/0741088318804820.

[5] M. Huerta and T. Garza, "Writing in Science: Why, How, and for Whom? A Systematic Literature Review of 20 Years of Intervention Research (1996–2016)," *Educational Psychology Review*, vol. 31, no. 3, pp. 533–570, Sep. 2019, doi: 10.1007/s10648-019-09477-1.

[6] E. B. D. Clabough and S. W. Clabough, "Using Rubrics as a Scientific Writing Instructional Method in Early Stage Undergraduate Neuroscience Study," *J Undergrad Neurosci Educ*, vol. 15, no. 1, pp. A85–A93, 2016.

[7] P. D. Klein and P. Boscolo, "Trends in research on writing as a learning activity," *Journal of Writing Research*, vol. 7, no. 3, pp. 311–350, 2016, doi: Journal of Writing Research.

[8] B. E. C. Timmerman, D. C. Strickland, R. L. Johnson, and J. R. Payne, "Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing," *Assessment & Evaluation in Higher Education*, vol. 36, no. 5, pp. 509–547, Aug. 2011, doi: 10.1080/02602930903540991.

[9] B. O'Donovan, C. Rust, and M. Price, "A scholarly approach to solving the feedback dilemma in practice," *Assessment & Evaluation in Higher Education*, vol. 41, no. 6, pp. 938–949, Aug. 2016, doi: 10.1080/02602938.2015.1052774.

[10] C. Evans, "Making Sense of Assessment Feedback in Higher Education," *Review of Educational Research*, vol. 83, no. 1, pp. 70–120, Mar. 2013, doi: 10.3102/0034654312474350.

[11] J. B. Garfield, D. Ben-Zvi, B. Chance, E. Medina, C. Roseth, and A. Zieffler, "Assessment in Statistics Education," in *Developing Students' Statistical Reasoning*, Dordrecht: Springer Netherlands, 2008, pp. 65–89. doi: 10.1007/978-1-4020-8383-9_4.

[12] A. Pisano, A. Crawford, H. Huffman, B. Graham, and N. Kelp, "Development and Validation of a Universal Science Writing Rubric That is Applicable to Diverse Genres of Science Writing," *J Microbiol Biol Educ.*, vol. 22, no. 3, pp. e00189-21, Dec. 2021, doi: 10.1128/jmbe.00189-21.

[13] V. Sampson, P. Enderle, J. Grooms, and S. Witte, "Writing to Learn by Learning to Write During the School Science Laboratory: Helping Middle and High School Students Develop Argumentative Writing Skills as They Learn Core Ideas," *Science Education*, vol. 97, no. 5, pp. 643–670, 2013, doi: 10.1002/sce.21069.

[14] J. Trace, V. Meier, and G. Janssen, "'I can see that': Developing shared rubric category interpretations through score negotiation," *Assessing Writing*, vol. 30, pp. 32–43, Oct. 2016, doi: 10.1016/j.asw.2016.08.001.

[15] T. H. Sundeen, "Instructional rubrics: Effects of presentation options on writing quality," *Assessing Writing*, vol. 21, pp. 74–88, Jul. 2014, doi: 10.1016/j.asw.2014.03.003.

[16] R. L. Bangert-Drowns, M. M. Hurley, and B. Wilkinson, "The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis," *Review of Educational Research*, vol. 74, no. 1, pp. 29–58, Mar. 2004, doi: 10.3102/00346543074001029.

[17] B. Hand, Y.-C. Chen, and J. K. Suh, "Does a Knowledge Generation Approach to Learning Benefit Students? A Systematic Review of Research on the Science Writing Heuristic Approach," *Educational Psychology Review*, Aug. 2020, doi: 10.1007/s10648-020-09550-0.

[18] B. Hand, M. C. Shelley, M. Laugerman, L. Fostvedt, and W. Therrien, "Improving critical thinking growth for disadvantaged groups within elementary school science: A randomized controlled trial using the Science Writing Heuristic approach," *Sci. Ed.*, vol. 102, no. 4, pp. 693–710, Jul. 2018, doi: 10.1002/sce.21341.

[19] N. S. Stephenson and N. P. Sadler-McKnight, "Developing critical thinking skills using the Science Writing Heuristic in the chemistry laboratory," *Chem. Educ. Res. Pract.*, vol. 17, no. 1, pp. 72–79, 2016, doi: 10.1039/C5RP00102A.

[20] M. M. Balgopal, A. M. A. Casper, A. M. Wallace, P. J. Laybourn, and E. Brisch, "Writing Matters: Writing-toLearn Activities Increase Undergraduate Performance in Cell Biology," *BioScience*, vol. 68, no. 6, pp. 445–454, Jun. 2018, doi: 10.1093/biosci/biy042.

[21] J. Airey and C. Linder, "A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes," *J. Res. Sci. Teach.*, vol. 46, no. 1, pp. 27–49, Jan. 2009, doi: 10.1002/tea.20265.

[22] P. Black and D. William, "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, Mar. 1998, doi: 10.1080/0969595980050102.

[23] C. S. Kalman, *Successful Science and Engineering Teaching: Theoretical and Learning Perspectives*. in Innovation and Change in Professional Education. Springer, Cham, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-66140-7

[24] United States Chamber of Commerce, "Bridging the soft skills gap: How the business and education sectors are partnering to prepare students for the 21 st century workforce," Center for Education and Workforce, Washington, DC, 2017. [Online]. Available: https://www.uschamberfoundation.org/sites/default/files/Closing%20the%20Soft%20Skills%20Gap.pdf

[25] National Science & Technology Council, "Charting a course for success: America's strategy for STEM education," Office of Science and Technology Policy, Washington, DC, 2018. [Online]. Available: https://www.energy.gov/sites/default/files/2019/05/f62/STEM-Education-Strategic-Plan-2018.pdf

[26] C. Zimmerman, "The development of scientific reasoning: What psychologists contribute to an understanding of elementary science learning," National Research Council Committee on Science Learning Kindergarten through Eighth Grade, 2005. [Online]. Available: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_080105.pdf

[27] L. Schauble, R. Glaser, R. A. Duschl, S. Schulze, and J. John, "Students' Understanding of the Objectives and Procedures of Experimentation in the Science Classroom," *Journal of the Learning Sciences*, vol. 4, no. 2, pp. 131–166, Apr. 1995, doi: 10.1207/s15327809jls0402_1.

[28] K. Hogan and M. Maglienti, "Comparing the epistemological underpinnings of students' and scientists' reasoning about conclusions," *J. Res. Sci. Teach.*, vol. 38, no. 6, pp. 663–687, Aug. 2001, doi: 10.1002/tea.1025.

[29] D. Kuhn, E. Amsel, M. O'Loughlin, L. Schauble, B. Leadbeater, and W. Yotive, *The development of scientific thinking skills*. Academic Press, 1988.

[30] F. Reif and J. H. Larkin, "Cognition in scientific and everyday domains: Comparison and learning implications," *J. Res. Sci. Teach.*, vol. 28, no. 9, pp. 733–760, Nov. 1991, doi: 10.1002/tea.3660280904.

[31] K. E. Stanovich and R. F. West, "Reasoning independently of prior belief and individual differences in actively open-minded thinking.," *Journal of Educational Psychology*, vol. 89, no. 2, pp. 342–357, Jun. 1997, doi: 10.1037/0022-0663.89.2.342.

[32] P. A. Klaczynski, D. H. Gordon, and J. Fauth, "Goal-oriented critical reasoning and individual differences in critical reasoning biases.," *Journal of Educational Psychology*, vol. 89, no. 3, pp. 470–485, Sep. 1997, doi: 10.1037/0022-0663.89.3.470.

[33] J. D. Bransford, A. L. Brown, and R. R. Cocking, *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, D.C.: National Academies Press, 2000, p. 9853. doi: 10.17226/9853.

[34] R. Bybee, B. McCrae, and R. Laurie, "PISA 2006: An assessment of scientific literacy," *J. Res. Sci. Teach.*, vol. 46, no. 8, pp. 865–883, Oct. 2009, doi: 10.1002/tea.20333.

[35] L. C. McDermott, P. S. Shaffer, and M. L. Rosenquist, *Physics by inquiry: an introduction to physics and the physical sciences*, vol. 1 and 2. New York: J. Wiley, 1996.

[36] D. R. Sokoloff, R. K. Thornton, and P. W. Laws, *RealTime physics: active learning laboratories, Modules 1-4*. New York: Wiley, 2004.

[37] L. Bao, K. Koenig, Y. Xiao, J. Fritchman, S. Zhou, and C. Chen, "Theoretical model and quantitative assessment of scientific thinking and reasoning," *Phys. Rev. Phys. Educ. Res.*, vol. 18, no. 1, p. 010115, Feb. 2022, doi: 10.1103/PhysRevPhysEducRes.18.010115.

[38] A. Zeineddin and F. Abd-El-Khalick, "Scientific reasoning and epistemological commitments: Coordination of theory and evidence among college science students," *J. Res. Sci. Teach.*, vol. 47, no. 9, pp. 1064–1093, Nov. 2010, doi: 10.1002/tea.20368.

[39] W. C. Sá, C. N. Kelley, C. Ho, and K. E. Stanovich, "Thinking about personal theories: individual differences in the coordination of theory and evidence," *Personality and Individual Differences*, vol. 38, no. 5, pp. 1149–1161, Apr. 2005, doi: 10.1016/j.paid.2004.07.012.

[40] S. L. Eddy and K. A. Hogan, "Getting Under the Hood: How and for Whom Does Increasing Course Structure Work?," *LSE*, vol. 13, no. 3, pp. 453–468, Sep. 2014, doi: 10.1187/cbe.14-03-0050.

[41] L. Tsui, "Effective Strategies to Increase Diversity in STEM Fields: A Review of the Research Literature," *Journal of Negro Education*, vol. 76, no. 4, pp. 555–581, 2007.

[42] D. Baldwin, M. Fowles, and S. Livingston, "Guidelines for constructed-responses and other performance assessmen," Educational Testing Service, Princeton, NJ, TOEFL BT 02 RR-07-02, 2008.

[43] E. D. Turley and C. W. Gallagher, "On the uses of rubrics: Reframing the great rubric debate," *The English Journal*, vol. 97, no. 4, pp. 87–92, 2008.

[44] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[45] M. J. Warrens, "Conditional inequalities between Cohen's kappa and weighted kappas," *Statistical Methodology*, vol. 10, no. 1, pp. 14–22, Jan. 2013, doi: 10.1016/j.stamet.2012.05.004.

[46] K. Krippendorff, *Content analysis: an introduction to its methodology*, Fourth Edition. Fourth Edition.Los Angeles: SAGE, 2018.

[47] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. the problems of two paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 543–549, Jan. 1990, doi: 10.1016/0895-4356(90)90158-L.

[48] D. V. Cicchetti and A. R. Feinstein, "High agreement but low kappa: II. Resolving the paradoxes," *Journal of Clinical Epidemiology*, vol. 43, no. 6, pp. 551–558, Jan. 1990, doi: 10.1016/0895-4356(90)90159-M.

[49] B. North, "Scales for rating language performance: Descriptive models, formulation styles, and presentation forma," Educational Testing Service, Princeton, NJ, TOEFL Monograph MS-24, 2003.

[50] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information.," *Psychological Review*, vol. 63, no. 2, pp. 81–97, Mar. 1956, doi: 10.1037/h0043158.

[51] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review*, vol. 2, no. 2, pp. 130–144, Jan. 2007, doi: 10.1016/j.edurev.2007.05.002.

[52] G. Janssen, V. Meier, and J. Trace, "Building a better rubric: Mixed methods rubric revision," *Assessing Writing*, vol. 26, pp. 51–66, Oct. 2015, doi: 10.1016/j.asw.2015.07.002.

[53] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[54] M. Gunel, B. Hand, and V. Prain, "Writing for Learning in Science: A Secondary Analysis of Six Studies," *Int J of Sci and Math Educ*, vol. 5, no. 4, pp. 615–637, Oct. 2007, doi: 10.1007/s10763-007-9082-y.

[55] A. Moon, A. R. Gere, and G. V. Shultz, "Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom," *Science Education*, vol. 102, no. 5, pp. 1007–1028, 2018, doi: 10.1002/sce.21454.