


Rubric for assessing thinking skills in free-response exam problems

Fatema Al-Salmani and Beth Thacker 

Department of Physics and Astronomy, Texas Tech University, MS 41051 Lubbock, 79409 Texas, USA



(Received 9 December 2020; accepted 20 April 2021; published 13 May 2021)

We designed a rubric to assess free-response exam problems in order to compare thinking skills evidenced in exams in classes taught by different pedagogies. The rubric was designed based on Bloom's taxonomy and then used to code exam problems. We have analyzed historical and recent exam problems in both algebra-based and calculus-based exams. In particular, we have examined cases where the same problem was administered on exams taught by different pedagogies. In some cases, we were able to compare different sections of the algebra-based physics course taught by the same instructor, sometimes during the same semester, one with inquiry-based instruction and the other in a more traditional lecture environment. We discuss the development of the instrument and present results. The inquiry-based students consistently demonstrated use of the thinking skills coded at least as often as the traditional students, and usually more often.

DOI: [10.1103/PhysRevPhysEducRes.17.010135](https://doi.org/10.1103/PhysRevPhysEducRes.17.010135)

I. INTRODUCTION

In recent years there has been a move toward recognition of the need to assess not just conceptual understanding, but also students' thinking skills. There has been significant work done by a number of groups both in the development of assessment instruments and on promoting thinking skills in the classroom [1–8]. Most of these authors define and discuss “critical thinking skills” in a particular context. We are interested in assessing the thinking skills employed by students in introductory physics courses as evidenced in free-response (FR) exam questions. In particular, we are interested in researching the responses of students taught by different pedagogies. We have previously analyzed the conceptual understanding of students taught by different pedagogies [9], and view this as a next step in studying the benefits and drawbacks of different instructional methods.

We define thinking skills broadly as the cognitive processes employed in answering a question, solving a problem, making a decision, analyzing facts, or creating new ideas. There already exists a widely used taxonomy of cognitive processes, Bloom's taxonomy [10–11], used as a framework by many for instruction and assessment. We decided to use this as the basis for a rubric designed to assess the cognitive skills employed by students in answering FR exam problems. Operationally, then, in our context,

thinking skills are defined by the “action descriptors” at each level of Bloom's taxonomy (revised version) [10].

We have historically administered locally written (both research-inspired and more traditional) free-response (FR) problems systematically across our introductory physics courses. The FR problems are used to assess mathematical, laboratory, and thinking skills used by students and to have a deeper understanding of the cognitive resources applied in problem solving. They were used as a complementary assessment to conceptual inventories. We observed differences in the FR assessment results across pedagogies [12] and decided that a more rigorous study of students' thinking skills is needed. This paper discusses that study.

We have studied both calculus-based and algebra-based classes. In this paper we focus on algebra-based classes. Some of the algebra-based course sections are taught in a laboratory-based, studio-style setting without a lecture and without a required text. Students work in groups on the evidence-based, laboratory-based materials [designed 15–20 years ago with National Science Foundation (NSF) funding [13–14]], with the instructors (faculty, graduate, and undergraduate teaching assistants) circulating about the room to guide the students' understanding through Socratic questioning. The materials are published and available to others [15]. As the learning is in an inquiry-based format, we will refer to the course as INQ. Originally, the INQ course was taught to class sizes of 24 students. Over the past three years, we have expanded the course to classes of 60 students, adding undergraduate learning assistants (LAs) with the expansion.

The comparison of the INQ course with the traditional sections, particularly as related to thinking skills, is of continued interest to us. While conceptual inventory scores in traditional sections are much lower than in the INQ

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

course, lecture sections taught with interactive engagement have similar conceptual inventory scores to the INQ course [9]. However, the responses to FR problems have led us to believe that the thinking skills applied to problem solving are different in the two populations.

As a first iteration, we designed a rubric based on Bloom's taxonomy (revised version) [10] for each problem analyzed. While we found more frequent use of all of the thinking skills observed in the students in the INQ classroom, we wanted to create a general rubric that we could use for all problems and could validate and test for reliability for use by others. Our goal was to design a simple, coarse rubric that would give a general idea of the level of thinking skills applied in exam problem solving. Once developed, we would use the rubric to answer the research question: How do the thinking skills of INQ and traditionally taught students compare as evidenced in the context of FR exam problems? This paper discusses the development of the rubric and presents the results of the rubric applied to sample FR questions on exams in INQ and more traditionally taught classes at our institution.

II. STUDENT POPULATIONS

A. Algebra-based course

The algebra-based introductory physics sequence consists of two semesters, the first covers mostly mechanics and the second mostly electricity and magnetism (E&M). The student population is predominantly health science majors, including premedical, pre dental, pre-physical therapy, etc. The number of students registered for the first semester is usually about 350 and around 150–200 students register for the second semester. Each semester, there are two INQ sections of 60 students each. The rest of the students are in lecture-based sections of 60–200 students depending on instructor schedules and classroom availability.

1. Lecture-based sections

The instruction in the lecture-based sections is primarily traditional lecture, although some instructors use interactive-engagement techniques. The traditional sections also have a one-hour recitation and a 2-h evidence-based lab each week. The labs and recitations are taught by graduate students and are common among the lecture instructors. Students from each of the lecture instructors are mixed in the labs and recitations.

2. INQ section

The INQ section was developed with National Science Foundation (NSF) funding 15–20 years ago and has been taught as one or more section(s) of the algebra-based course every semester since then. It was developed explicitly for health science majors, taking their needs, learning styles, backgrounds and motivations into account. It is taught

without a required textbook in a studio-style environment with Socratic questioning pedagogy. Students work through the units in groups, learning to develop both quantitative and qualitative models based on their observations and inferences and then using the models to make predictions and solve problems. The goal is to start in the laboratory like scientists and learn the content through (guided) experimentation. The materials consist of the laboratory units, pretests, readings, and exercises. There are also homework sets, exams, and quizzes. The course covers approximately the same content as is covered in the other sections of the class, but with more of a focus on developing models based on experimentation and on developing observational, analytic, and thinking skills. Students can choose to be in either the INQ or traditional section, but most often it is determined by what fits their schedules and when they need to take the course, as far as graduation. We do not see (and have not seen historically) any difference in conceptual inventory pretests between the traditional and INQ sections. [For example, students usually score around 20% correct on the Brief Electricity and Magnetism Assessment [16] as a pretest. In 2019, the pretest data were 19.1 ± 0.9 (INQ), 20.1 ± 0.9 (traditional). Historical data can be found in Ref. [9].]

III. METHODS

A. Identification of common exam problems

We have a pool of exam problems that have been administered across multiple courses, multiple sections of courses, and across multiple years. We identified problems that were common across different sections of the algebra-based course. We also identified sections of the course taught by the same instructor using different pedagogies. (Some of the INQ instructors also teach lecture sections, using either traditional or interactive-engagement strategies.) We identified seven problems that had been administered across the algebra-based sections that include at least one INQ section and at least one traditional or interactive-engagement lecture section. From this pool, we chose two on the same topic, one research inspired, one more traditional (not interactive engagement), to establish interrater reliability.

B. The rubric and analysis

1. The rubric

As stated above, we first designed a rubric for each individual problem, as our first attempts at a general rubric were not successful. We did learn from that experience, though. In particular, we found that we needed to add a zeroth level to Bloom's taxonomy that included answers that were blank, totally incorrect, just statements or partially remembered facts or concepts. We also observed that we had evidence mostly of the three lowest levels of Bloom's taxonomy and not often the three higher-level

The Rubric	
Bloom's taxonomy levels for physics free-response exam problem solutions:	
0) i) Nothing	
ii) Totally incorrect	
iii) Statement only	
iv) Partial Remember – includes some facts or concepts needed to solve a problem, but at least one necessary concept or fact is missing or incorrectly remembered.	
1) Remember: Recall facts and basic concepts	Solution includes recollection of a fact or concept or set of facts or concepts needed to solve the problem in either mathematical or written form. (If one or more necessary concepts are missing, it counts as Partial Remember and not Remember.)
2) Understand: Explain ideas or concepts	Solution includes a correct application of the facts or concepts recalled. This could be verbal or mathematical; it includes recognition of any details or constraints given in the particular problem and how they would be applied. A picture must be supported by a verbal explanation.
3) Apply/Analyze: Use information in new situations/draw connections among ideas	Solution demonstrates a valid, logical and consistent process for determining an answer, applying (1) and (2) and demonstrates evidence of understanding an underlying model or fundamental principle.
4) Evaluate: Justify a stand or decision	Solution includes a comparison of or discrimination between different possibilities or an explanation of why an answer is consistent or makes sense or why it is incorrect.

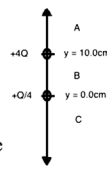
FIG. 1. The rubric.

thinking skills. As these were existing exam problems, this was actually not too surprising. Most of the problems were not written to explicitly elicit higher-level thinking skills.

In the development of the general rubric, we started with a rubric with seven levels (Bloom's six levels plus a zeroth level) but soon moved to a five-level rubric (zeroth level plus four levels). We did not observe the level *create* and had trouble distinguishing the levels *apply* and *analyze* in our context. So, our rubric has five levels (*zeroth, remember, understand, apply and analyze, evaluate*) as shown in Fig. 1. The rubric was designed so that each student's answer would be coded as to whether or not the written answer supplied evidence of the thinking skill described at a particular level of the rubric. The student's answer was coded with a 1 or a 0 for each level on each part of the problem (for multipart problems).

A sample problem is shown in Fig. 2. In Fig. 3, we give examples of how each level is coded with explanations of why it was coded that way.

A positive charge $+4Q$ is placed at $y = 10.0\text{cm}$ and a charge of $+Q/4$ is placed at the origin of the y -axis, as in the picture below. Three regions of the y -axis are labeled. Region A is $y > 10.0\text{cm}$. Region B is $0.0\text{cm} < y < 10.0\text{cm}$. Region C is $y < 0.0\text{cm}$.



a) Is there any position along the y -axis that the net electric field could be zero? If so, in which region(s) could the net electric field be zero. Explain your reasoning.

b) If there is a position along the y -axis at which the net electric field is zero, determine that location. If there is not a position where the net electric field could be zero, explain why not. Show your work and explain your reasoning.

FIG. 2. A problem administered to two algebra-based physics sections, one INQ, one traditional, both taught by the same instructor, the same semester, different pedagogies.

2. The analysis

For the development of the general rubric, we first chose two problems shown in Figs. 4 and 5, one a traditional exam problem (Fig. 4) [17] and the other a research-inspired problem (Fig. 5) [18–19]. We worked first with a subset of the 160 student answers. We chose 10 INQ and 10 traditional student answers and each rater coded each problem. The student answers were chosen randomly. While we know that the student population in the algebra-based course is approximately 50% female and 50% male and approximately 30% Hispanic each semester, we do not have other demographic information and we were not studying demographic results. The raters did not know if the answers were from traditional or INQ students. We met to compare our coding and discuss our agreements, disagreements and any adjustments that needed to be made to the rubric. We then coded another subset of the two problems and compared our answers. After achieving 90% agreement on these two problems, we began the analysis of the other problems.

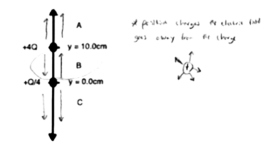
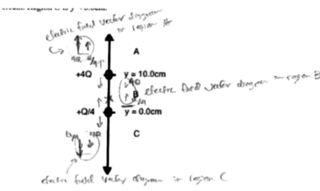
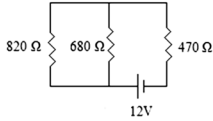
None	$E = \frac{kQ}{r^2}$ <p>The net electric field could be zero in region B because 2 positive charges will repel each other so between the two is where the net electric field is 0.</p>	Students answer include statement only
Remember	Where the electric field from the two charges are equal and opposite the electric field will be zero. This point lies in area B.	Student answers include set of concepts needed to solve the question
Understand	The net force could be zero in region B. When viewing regions, A and C the forces are in the same direction. However, you are wanting a balance of forces. One finds region B to have a balance of forces by finding the distance and charges of $-Q$.	Students answer include correct application of concepts and how they would be applied.
Apply/Analyze	<p>Yes, there is a position along y axis where the net electric field is zero and the region is Region B. It's in region B because the electric fields from both the charges are pointing in the opposite direction causing them at some point to equal each other and making the net electric field to be zero.</p> 	The solution demonstrates logical and consistent process to determine the answer. The picture is supporting the student wording and demonstrates evidence of understanding
Evaluate	<p>Yes, Region B is where the $\Sigma E=0$. Each charge is (+) charge which because the electric field E points radially outward from the point charge. The only region where the directions of the electric field vector from each point are in opposite directions and thus cancel each other is region B. In addition, the position where $\Sigma E=0$ is closer to the charge of less magnitude i.e. $+Q/4$, so that the overall sum can equal 0 (common sense).</p> 	Student answer includes an explanation of why the answer is making sense.

FIG. 3. Examples of students answers for part (a) of the sample problem.

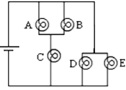
• Consider the circuit shown in the diagram below.



a) Determine the total resistance of the circuit. Show your work.
 b) Calculate the potential difference across the 680 Ω resistor. Show your work.
 c) Calculate the current through the 820 Ω resistor. Show your work.

FIG. 4. Traditional problem used in adjusting rubric.

1) Consider the following circuit. All of the bulbs are identical.



a) Rank the brightness of (current through) the bulbs from greatest to least. Explain why you ranked the bulbs the way you did.
 b) If bulb B were unscrewed from the circuit, would the current through
 i) bulb C increase, decrease or remain the same? Explain your reasoning.
 ii) bulb D increase, decrease or remain the same? Explain your reasoning.
 c) Bulb B is returned to the circuit. If a bulb were added in parallel to bulb B, would the potential difference across B increase, decrease or remain the same? Explain your reasoning.

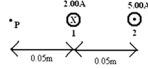
FIG. 5. Research-inspired problem used in adjusting rubric.

We also interviewed 13 students in Spring 2019 on one problem to establish construct validity.

We looked for agreement between the students’ written and verbal answers. Disagreement (things added or left out of the video compared to the written) was extremely rare in the verbal versions of the students’ responses. In the interviews, we did push students on their understanding of some concepts beyond the written questions, but where they discussed their written answers, what they had written and why, the written and verbal answers were in agreement. The verbal discussion may have included more words, or in some cases rethinking, but the verbal answers were consistent with the students’ written answers.

We report the results of two sample problems. The first, shown in Fig. 2, was administered in two sections the same semester (Spring 2017) taught by the same instructor with two different pedagogies. One section was taught in the INQ format. The other was a traditional (noninteractive-engagement) lecture-based section. The class sizes were approximately the same, with the INQ section having 53 students and the traditional section 57 students. The second example problem shown in Fig. 6, was administered to four different populations all taught by the same instructor: (i) a small INQ class in 2010 (24 students), (ii) a larger interactive-engagement class (80 students) the same 2010 semester, (iii) a larger INQ class in 2017 (60 students), (iv) a larger INQ class in 2019 (60 students). We report Cohen’s kappa for the two raters for interrater reliability and the results of Fisher’s exact test to determine

• Consider two parallel, current-carrying wires, as in the figure below. Wire 1 carries 2.00A of current into the page. Wire 2 carries 5.00A of current out of the page. The wires are 0.05m apart.



a) Determine the magnitude and direction of the magnetic field at point P. Show your work.
 b) A third wire carrying current $I = 1.00\text{A}$ into the page, is placed at point P. Determine the magnitude and direction of the net force on the third wire, if the wire is 1.00m long. Show your work

FIG. 6. A problem administered to four algebra-based physics sections, one small (24 students) INQ in 2010, one traditional interactive engagement (80 students) in 2010, two larger (60 students) INQ sections in 2017 and 2019. All classes had the same instructor.

if there is a significant difference between the courses based on our results in Sec. IV.

IV. RESULTS

A. Interrater reliability

After establishing interrater reliability, all of the problems were coded by the same two raters. We also introduced the rubric to two additional local raters and had them code some of the problems. The weighted Cohen’s kappa was greater than 0.80 when any two of the raters were compared.

We did not, on any of the problems, assume the rubric levels were hierarchical (if a level is coded with a 1, all of the levels below it are also coded 1), but we observed it for all of the problems analyzed so far. For this reason, the weighted Cohen’s kappa is the more relevant statistic. However, we report both the weighted and unweighted results for each problem.

For the problem in Fig. 2, the unweighted Cohen’s kappa results were 0.70 and 0.94 and the weighted Cohen’s kappa results were 0.86 and 0.97 on the (a) and (b) parts of the problem for the traditional class. For the INQ class, the unweighted Cohen’s kappa results were 0.68 and 0.71 and the weighted Cohen’s kappa results were 0.82 and 0.85 on parts (a) and (b), respectively.

For the problem in Fig. 6, the unweighted Cohen’s kappa results were all around 0.70 or higher and the weighted Cohen’s kappa results were 0.80 or higher on the (a) and (b) parts of the problem for each of the sections analyzed.

B. Comparison results

In Figs. 7 and 8, we show the results of the comparison of students taught by two different pedagogies, same semester, same instructor. The traditional class was taught very traditionally, not using interactive engagement. This is compared to the INQ class. Figure 7 shows the results of part (a) and Fig. 8 the results of part (b) of the problem shown in Fig. 2. A much higher percentage of the INQ students demonstrated each of the thinking skills assessed by our rubric. This is consistent with our previous data

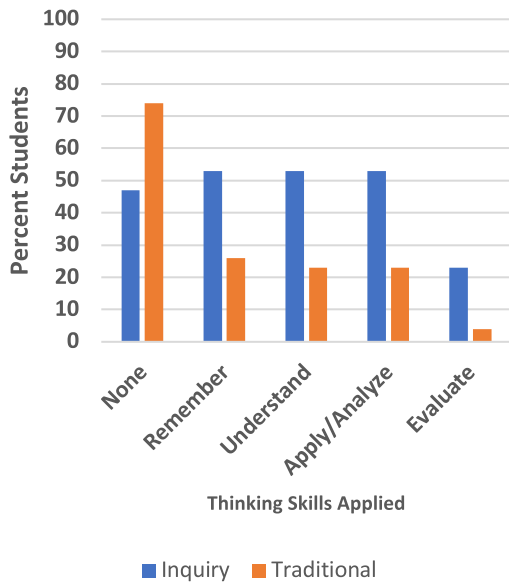


FIG. 7. The results of the general rubric applied to two algebra-based physics sections, one INQ, one traditional, part (a). Same instructor, different pedagogies.

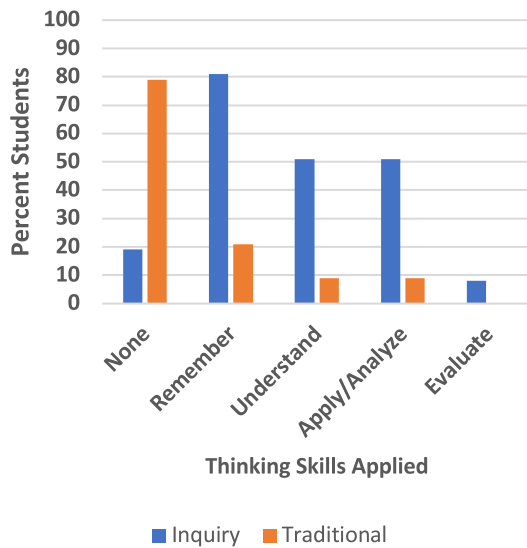


FIG. 8. The results of the general rubric applied to two algebra-based physics sections, one INQ, one more traditional, part (b). Same instructor, different pedagogies.

using a problem specific rubric. It is also an example of the difference in thinking skills demonstrated on an exam problem by INQ and traditional students. We have evidence of much more frequent use of all of the thinking skills by the INQ students.

The results of Fisher’s exact test were $p = 0.0014$ and $p = 0.00001$ for part (a) and (b), respectively, indicating that the results from the two classes were significantly different.

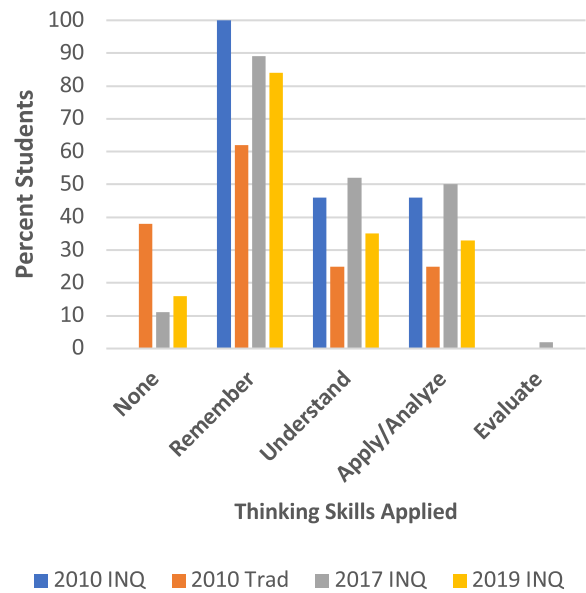


FIG. 9. The results of the general rubric applied to four algebra-based physics sections, one small (24 students) INQ, one traditional interactive engagement (80 students) in 2010, two larger INQ (60 students) in 2017 and 2019, part (a). Same instructor for all sections.

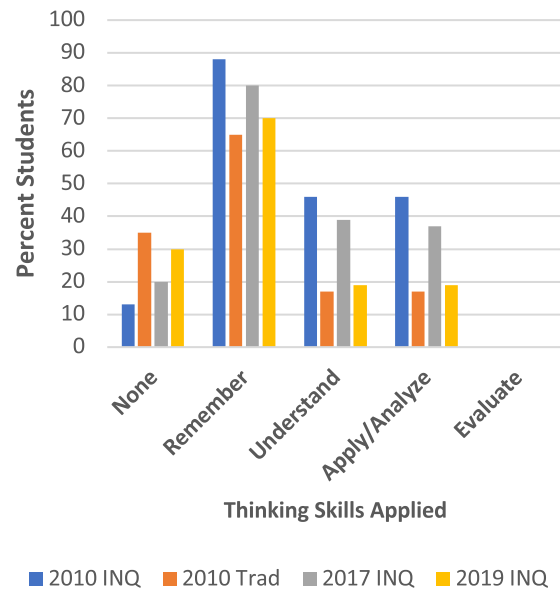


FIG. 10. The results of the general rubric applied to four algebra-based physics sections, one small (24 students) INQ, one traditional interactive engagement (80 students) in 2010, two larger INQ (60 students) in 2017 and 2019, part (b). Same instructor for all sections.

In Figs. 9 and 10, we show the results of the problem shown in Fig. 6 for the same professor teaching both INQ and traditional with interactive engagement in 2010 and the same professor, same problem, with larger INQ sections in 2017 and 2019.

We compared the results using Fisher's exact test. The 2010 and 2017 INQ classes were significantly different from the 2010 traditional class. Comparing the 2010 INQ and traditional classes in part(a), we found $p = 0.0003$ and in part (b), $p = 0.0062$.

Comparing the 2017 INQ and the traditional class, in part (a), $p = 0.0007$, and in part (b), $p = 0.020$. The 2019 INQ class was not significantly different from the traditional class. The only other significant difference was in part (a) of the comparison of the 2019 and 2010 INQ classes, $p = 0.0211$.

V. DISCUSSION AND CONCLUSIONS

We have presented a description of the development process and two examples of the application of a general rubric for assessing thinking skills based on Bloom's taxonomy. In the first example, we compared two sections taught by the same instructor the same semester. One section was inquiry based, the other traditional. The INQ section demonstrated more frequent use of all of the thinking skills analyzed. Also, some of the INQ students were coded with "evaluate," even though that was not explicitly asked for in the problem. This did not happen in the traditional section. While this is only one example, we observed the INQ students demonstrating the thinking skills coded more often than traditionally taught students in almost all of the problems analyzed.

In the second problem, we have presented data from the same instructor teaching a smaller (24 students) INQ section and an interactive-engagement traditional section (80 students) in 2010. In the same graph we presented the same instructor in two larger INQ sections in 2017 and 2019. This again shows an INQ and traditional comparison, with the INQ students demonstrating the thinking skills coded at least as often and usually more often than the traditional students. It also shows (a) some decrease in the

application of thinking skills with the expansion of the INQ sections and (b) that the thinking skills of the expanded INQ sections can vary significantly semester by semester. Still the INQ sections demonstrated at least the same level of use of thinking skills coded, and usually higher than the traditional section.

In conclusion, we have developed and demonstrated the use of a general rubric for analyzing thinking skills at our institution. We have used the instrument to compare the thinking skills of students taught by different pedagogies in the context of FR exam problems and to analyze changes in thinking skills applied by students in the INQ class with expansion of INQ class size.

We plan to continue this research at our own institution to study the thinking skills employed by students taught with different pedagogies in the context of FR exam problems. We will also expand our project locally to (i) the analysis of FR homework problems, (ii) the development of homework and exam problems designed to explicitly elicit higher order (as defined by our rubric) thinking skills, and (iii) comparison to existing, valid and reliable noncontent specific thinking skills assessment instruments, such as the Critical Thinking Assessment Test (CAT) or California Critical Thinking Skills Test (CCTST) [20–21]. We also plan to expand the study to other universities, making the instrument available to researchers and instructors at other institutions to further validate and establish reliability and to introduce them to a relatively easy to use instrument for research and assessment of thinking skills.

ACKNOWLEDGMENTS

We acknowledge the contributions of Nirosha Kandedgara, Riley Dubbels, and Victor Principe for helping with rubric development and/or interrater reliability.

-
- [1] N. Holmes and C. Weiman, Introductory physics labs: We can do better, *Phys. Today* **71**, No. 1, 38 (2018).
 - [2] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Phys. Rev. Phys. Educ. Res.* **15**, 010135 (2019).
 - [3] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11199 (2015).
 - [4] E. Etkina and G. Planinsic, Defining and developing "critical thinking" through devising and testing multiple explanations of the same phenomenon, *Phys. Teach.* **53**, 432 (2015).
 - [5] D. Tiruneh, M. De Cock, A. Weldeslassie, J. Elen, and R. Janssen, Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism, *Int. J. Sci. Math. Educ.* **15**, 663 (2017).
 - [6] I. Puspita, I. Kaniawati, and I. R. Suwarma, Analysis of critical thinking skills on the topic of static fluid, *J. Phys. Conf. Ser.* **895**, 012100 (2017).
 - [7] W. Wartono, M. Nur Hudha, and J. Batiolona, How are the physics critical thinking skills of the students taught by using inquiry-discovery through emperical theoretical overview, *EURASIA J. Math. Sci. Tech. Educ.* **14**, 691 (2017).

- [8] T. Sugiarti, I. Kaniawati, and L. Aviyanti, Development of assessment instrument of critical thinking in physics at senior high school, *J. Phys. Conf. Ser.* **812**, 012018 (2017).
- [9] B. Thacker, H. Dulli, D. Pattillo, and K. West, Lessons from a large-scale assessment: Results from conceptual inventories, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020104 (2014).
- [10] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, New York, 2001).
- [11] B. S. Bloom, *Taxonomy of Educational Objectives: The Classification of Educational Goals* (Longmans, Green, New York, NY, 1956).
- [12] B. Thacker, G. Chapagain, V. Rusuriye, and K. West, Beyond conceptual inventories: Lessons from a large-scale assessment, *US-China Educ. Rev. A* **7**, 1 (2017).
- [13] National Science Foundation, Course, Curriculum and Laboratory Improvement CCLI #9981031, in *Workshop Physics with Health Science Applications, 2000–2003*.
- [14] National Science Foundation, Course, Curriculum and Laboratory Improvement grant CCLI-EMD #0088780, in *Humanized Physics—Reforming Physics Using Multimedia and Mathematical Modeling, 2001–2004*.
- [15] B. Thacker and K. West, *Inquiry-Based Experimental Physics Volumes I and II* (Stipes Publishing L.L.C., Illinois, 2019).
- [16] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief Electricity and Magnetism Assessment, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010105 (2006).
- [17] D. Giancoli, *Physics Principles with Applications* (Pearson, Upper Saddle River, NJ, 2005).
- [18] L. McDermott and P. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding, *Am. J. Phys.* **60**, 994 (1992).
- [19] L. McDermott and P. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part II: Design of instructional strategies, *Am. J. Phys.* **60**, 1003 (1992).
- [20] B. Stein, A. Haynes, M. Redding, T. Ennis, and M. Cecil, Assessing critical thinking in STEM and beyond, in *Innovations in E-learning, Instruction Technology, Assessment, and Engineering Education*, edited by M. Iskander (Springer, New York, 2007), pp. 79–82.
- [21] P. A. Facione, Using the California Critical Thinking Skills Test in research, evaluation, and assessment (ERIC Document Reproduction Service No. ED 337 498, 1991).