How the introduction of self-assessment rubrics helped students and teachers in a project laboratory course

Sergej Faletič* and Gorazd Planinšič

University of Ljubljana, Faculty of Mathematics and Physics, Jadranska ulica 19, 1000 Ljubljana, Slovenia

(Received 28 November 2019; accepted 21 October 2020; published 20 November 2020)

Scientific abilities rubrics developed at Rutgers University are an assessment and self-assessment tool. They consist of tables, each representing a broad scientific ability (for example, "to collect and analyze data"), the listing of subabilities (for example, "independent and dependent variables are identified") and the criteria to assess to which degree the subability has been developed. We introduced scientific abilities rubrics to assess student work in the *project laboratory*, a project-based course where groups of students solve open-ended experimental physics problems. They submit a report in the form of a web page, which is evaluated and returned to the students with feedback on what needs to be improved. These iterations are repeated until the report is deemed acceptable. We present our experience with the rubrics, and the development of a new rubric for the web report. We show that the introduction of the rubrics reduced the workload of the graders, while increasing the quality of the reports. Based on these results we conclude that using scientific abilities rubrics is a very efficient way of assessing experimental project-based work. We also present an analysis of which abilities, assessed by the rubrics, appear to be the most difficult to develop.

DOI: 10.1103/PhysRevPhysEducRes.16.020136

I. INTRODUCTION

A. Motivation and research questions

Project work is an increasingly important instructional method. It helps students develop the skills needed to design, run, and complete an experimental project, as well as to present its results. Physics curricula have a long tradition of so-called traditional laboratories, where the student is guided through the process of completing a particular experimental procedure using a set of step-by-step instructions. For simplicity we will use in this article the terms *cookbook* and *confirmational* experiments. The term cookbook experiments refers to experiments that give students very little freedom, as most of what they need to do is explicitly stated, sometimes down to the point of exactly which values to use in the experiment. The term confirmational experiments refers to experiments where the theoretical model is given in advance and the purpose of the experiment is to confirm (validate) the model experimentally. We want to make an explicit distinction between a confirmational and a *testing experiment* [1]. A testing experiment is designed to reject a given hypothesis, while a confirmational experiment is designed to confirm it. Sometimes instructions accompanying confirmational experiments go as far as specifying exactly what values of controllable variables to use, to avoid ending up outside the range of validity of the model. The laboratories that contain such experiments are not without merit. They usually focus on familiarizing the students with equipment and occasionally with additional or more specific content. Research shows, however, that traditional laboratory work does not significantly help the students improve their content knowledge [2], nor does it help develop research abilities needed to design and complete a research project [3–6].

Project-based work has been shown to develop these abilities more efficiently [7–9]. However, students' project work is difficult to assess. Assessment strategies employed in inquiry-based learning in general include assessment of work during the project, assessment of reports, and assessment of project outcomes [10]. Assessment tools involve questionnaires, tests, rubrics, and written feedback.

In the course *project laboratory* at the department of physics of the faculty of mathematics and physics, University of Ljubljana, we assess project work via reports that the students submit after the project work is completed. Graders evaluate the reports and provide feedback to the students to improve them. Previously, this feedback was in the form of written comments, and was a very time consuming practice. In 2014 we decided to try using the Rutgers scientific ability rubrics for assessment [11–13]. Following the introduction of the rubrics, we assessed their

*sergej.faletic@fmf.uni-lj.si

effects on the project laboratory course. We posed the following research questions:

(i) To what extent do scientific abilities rubrics help reduce the grader's workload in the project laboratory course?
(ii) How do scientific abilities rubrics affect the quality of students' work in open-ended project laboratory tasks?
(iii) How quickly do students develop certain abilities and can the scientific abilities rubrics be used to evaluate this?

For simplicity, we will refer to the period before the introduction of the rubrics as the *prerubrics period*, and the period after their introduction as the *rubrics period*. We will refer to the fall semester of 2014, during which we introduced a shortened version of the rubrics, as the *intermediate period*.

The structure of this paper is as follows: In Sec. I we give a literature overview on each research question. In Sec. II we discuss the setting. In Sec. III we discuss the instruments for data collection and the methods of analysis for each research question. In Sec. IV we present the findings related to each research question. In Sec. V we discuss the findings for each research question and in Sec. VI we present our answers to the research questions. The research questions are identified as *grader workload*, *quality of the reports,* and *developing scientific abilities*.

B. Review of the literature

1. Scientific abilities rubrics

A rubric is a set of guidelines that defines expectations related to the learning goals in a given course. Rubrics often come in the form of a table where specific scores are assigned to different levels of the achievement of the learning goals. For the Rutgers scientific abilities rubrics, these goals are proficiencies in the processes and procedures in which physicists engage when developing new knowledge, or apply it [12,13]. One line from such a rubric is given in Table I and an entire rubric is presented in the Appendix. The levels are labeled: "missing," indicating that the use of the ability has not been observed; "inadequate," indicating major flaws or omissions in the use of the ability; "needs improvement," indicating that the ability is used mostly adequately, but still requires small improvements; and "adequate," indicating an adequately developed and used ability. For convenience, the levels are assigned numerical scores 0, 1, 2, and 3, respectively. The goals are consistent with the general learning goals of the *investigative science learning environment (ISLE)* [1,13,14]. The term *scientific abilities* describes some of the most important procedures, processes, and methods that scientists use when constructing knowledge and when solving experimental problems. Etkina and colleagues used the term scientific abilities instead of science process skills, because the term "skills" might imply something that can be automatized through repetition, like riding a bicycle, while these abilities need to be used reflectively and critically.

The Rutgers group identified seven abilities. These are (labeled as in the rubrics in Ref. [11]) the abilities to (A) "represent information in multiple ways," (B) "design and conduct an experiment to investigate a phenomenon," (C) "design and conduct a testing experiment (testing an idea, hypothesis, explanation, or mathematical relation)," (D) "design and conduct an application experiment," (F) "communicate scientific ideas," (G) "collect and analyze experimental data," and (I) "evaluate models, equations, solutions, and claims." Furthermore, there are over 30 subabilities for which assessment rubrics have been developed (for more information see Refs. [1,12,13]). The rubric for ability B, to design and conduct an experiment to investigate a phenomenon, and its subabilities is shown in Fig. 7 in the Appendix; all other rubrics can be found at the website in Ref. [11].

Studies conducted by the Rutgers group found that the students who use rubrics for self-assessment of their experimental work transfer these abilities to new content and different contexts [15]. It takes a significant time for the majority of students to develop proficiency in using the abilities when designing, conducting, analyzing, and reporting experiments. However, after an average of 25–30 h of practicing, the majority of the students develop them at an acceptable level [16,17].

2. Grader workload

In the field of economics education, McGoldrick and Peterson [18] report that the rubrics can reduce the time

TABLE I. An example of a line in a rubric. This example is taken from the rubric to assess the "ability to collect and analyze experimental data."

	Ability	Missing	Inadequate	Needs improvement	Adequate
G4	Is able to record and represent data in a meaningful way	Data are either absent or incomprehensible.	Some important data are absent or incomprehensible. They are not organized in tables or the tables are not labeled properly.	All important data are present, but recorded in a way that requires some effort to comprehend. The tables are labeled but labels are confusing.	All important data are present, organized, and recorded clearly. The tables are labeled and placed in a logical order.

spent assessing learning as well as the subjectivity in grading, while also making the class activities more in line with the learning objectives. They do not, however, quantify any of these benefits. In physics education, an early form similar to a rubric is reported by Allie *et al.* [19]. Upon introduction, teachers expected an increase in grading time, but noticed none. A decrease in grading time is not reported. Holmes and Wieman [6] suggest that ISLE-based labs [12] and SQILabs [20] use group reports and rubrics to reduce grading time, but the reduction is not quantified in either reference.

3. Student performance

Other studies address the effectiveness of open-ended laboratories and the use of rubrics. Numerous reports find that open-ended laboratory work enhances students' performance in a generic way [3–5]. Strubbe and colleagues [21] investigated the effect of traditional laboratory work and SQILabs [20] on students' beliefs about physics, using the E-CLASS tool [22]. They report that after traditional laboratories students' beliefs are less expertlike than before the laboratories, while open-ended SQILabs laboratories change their beliefs towards more expertlike. Wilcox and Lewandowski [22] arrived at similar results investigating the effect of traditional and open-ended laboratories using the USA national data set of results of the E-CLASS survey. Berg *et al.* [23] compared an open-ended and a traditional version of the same experiment and found that the students engaged in the open-ended experiment had significantly higher learning gains than the students engaged in the traditional experiment at higher taxonomy levels (application, analysis, synthesis, and evaluation) Holmes and Wieman [6] specifically mention ISLE-based laboratories [12] and SQILabs [20] as examples of open-ended laboratories, which develop students' research abilities. Both of these use rubrics to provide feedback and guidance.

The efficiency of rubrics in providing quality feedback extends beyond laboratory work. Mason and Singh [24] report a study, where students were given an identical subset of problems in midterm and final exams, and despite being given the answers to midterm exams, their answers to the same problems in the final exam did not improve. This suggests that students would benefit from a more specific focus on self-reflection, and guidance on what and how to improve. Research suggests that rubrics can provide such support [25–28]. Ene and Kosobucki [29] report in a case study that the use of rubrics on its own improved a student's essay writing, although more concrete feedback comments were valued more. Cockett and Jackson [30] did a metastudy on the use of rubrics in nursing education. They found that the use of rubrics increases self-assessment and self-regulation of students and the transparency of grading criteria. The largest increase was observed when students cocreated the rubrics. Howell [31] found that the use of rubrics was the strongest predictor of students' achievement on an apply-theory-to-practice assignment in a juvenile delinquency class, with its effect size 0.49 nearly the double of the second predictor. In a metastudy, Jonsson and Svingby [28] conclude that rubrics improve transparency, facilitate teacher's feedback to students, and increase students' self-assessment. And, while it is not possible to say that the use of rubrics alone improves learning, because they are mostly combined with other interventions, the majority of the studies show improvement at least in some areas, for example, on essay type questions, but not on multiple choice questions, or in biology and algebra, but not in english and government [32]. In areas where an improvement was not observed, no negative effects were observed. In another metastudy, Panadero and Jonsson [26] arrived at similar conclusions, adding that rubrics decrease anxiety in students before exams, because of the increased transparency of the expectations. The primary cause of the beneficial effects on students appears to be the increased self-assessment [33].

4. Developing scientific abilities

Rutgers scientific abilities rubrics were purposefully designed to address scientific abilities, rather than specific content. A study by Rajapaksha and Hirsch [34] show that competency-based teaching increases learning outcomes compared to content-oriented teaching. A previous study using the Rutgers scientific abilities rubrics [16] showed which abilities are the most difficult to develop. However, our setting is different and thus provides additional information on the topic. In Etkina and colleagues' study [16] they examined a series of different laboratory activities. For each activity the students handed in individual reports. The reports were scored and the feedback returned to the students in the form of scores on the rubric. Then the process was repeated for the next laboratory, and so on. The development of each student's abilities was followed by the researchers throughout the course. Because of the changing of experiments, this study included the transfer of the abilities between various contexts. In our setting, the students complete one project and hand in one report per group, which is then scored and the feedback returned to the students mostly in the form of scores on the rubrics, but also with short comments on more specific issues. The students then improve the report and hand it in again. The process continues until the report receives adequate scores on all the rubrics. The students can improve a specific report multiple times. Transfer between contexts could not be investigated, because each group worked on only one project.

II. STUDY SETTING

Project laboratory is an elective course at the department of physics, faculty of mathematics and physics, University of Ljubljana offered to first and second year physics students [35]. The course is based on open-ended tasks that help students develop scientific abilities [12]. The

focus on scientific abilities is in line with the educational recommendations in Europe and the U.S. [36–38] and the choice of open-ended problems is in line with the research data, which shows that traditional laboratory work is inefficient in developing these abilities [3–6].

Students in the project laboratory course work on open-ended experimental physics problems. They work in groups of four to five students, which they form by themselves. The course is structured in the following way: in week 1 students are assigned a problem, they study it and come up with ideas on how to solve it. In week 2 they refine their ideas after discussion with the grader. In weeks 3–5 students perform experimental work in the laboratory for 3 h per week. Problems are of one of the three types (classification was adopted from Ref. [39]): *observational experiments*, *testing experiments*, and *application experiments*. In observational experiments, students are expected to collect data, identify patterns and come up with an explanation and a model for an observed phenomenon. In testing experiments, the students are given two or more hypotheses that attempt to explain the same phenomenon, and have to design and conduct experiments to differentiate between those hypotheses. This is done by rejecting or failing to reject each hypothesis. The best testing experiments are those that can reject multiple hypotheses. In application experiments, the students have to use previously developed knowledge to solve a practical problem. In weeks 6 and 7 students write the report on their research in the form of a web page. This report is assessed by the grader and returned to the students to be revised and improved. In weeks 8–15 students improve their reports based on the feedback provided by the grader. Students are given one week for improvement at each iteration. The iterations continue until the report is considered acceptable, but not longer than until the end of the semester. The goal is that the reports receive an adequate score on all subabilities. The groups that finish early are not given new assignments, because the goal of the course has been achieved.

The reports are graded "pass" or "fail." The criterion for a passing grade in the prerubrics period was how well a report matched the form and content of scientific papers, based on the grader's personal experience. It was the grader's professional decision whether a report was graded pass or fail. In the rubrics period the criterion is receiving an adequate score on all rubrics. In the prerubrics period the pass rate was 85% ($N = 33$) and in the rubrics period it was 95% ($N = 22$), considering only the reports analyzed in this paper.

Prior to the introduction of the rubrics in the course, the students had a short lesson at the beginning of the semester, where the course instructor explained what was expected in terms of the research and the report, as well as commented on some issues common in previous reports. Since the introduction of the rubrics, this step has been replaced with a short lesson on how to use the rubrics. Students receive the rubrics at the beginning of the project and can use them to guide their work throughout the course.

III. METHODS

A. The grading in the prerubrics and rubrics periods

Assessment of students' abilities in the project laboratory course is based on the written reports of the groups. Each group submits one report, which is assessed by the teaching assistant. Before the introduction of the rubrics, the assessment was entirely subjective. The two or three teaching assistants involved in the course each year agreed on the basic goals that have to be assessed: (i) a clear description of the investigation process, (ii) labels on figures and graphs, (iii) argumentation of claims, (iv) correctness of the physics, and (v) correctness of the data analysis, including attention to experimental uncertainties. Teaching assistants commented on the report and returned it to the students for improvement. Students then typically improved the report and turned it in again for a second evaluation, and the process continued until the report was satisfactory or the students gave up, which happened in 5 out of 33 reports in the prerubrics period and 1 out of 22 reports in the rubrics period.

In the fall semester of 2014 (referred to as the intermediate period), we introduced a shortened version of the Rutgers scientific abilities rubrics. Among the Rutgers rubrics, we selected two, because we were afraid that introducing more rubrics would be overwhelming for the students who have never seen rubrics before. We chose rubric B for the ability to design and conduct an experiment to investigate a phenomenon, and rubric F for the ability to communicate scientific ideas. We expected the feedback to the students to consist only of the scores on the rubrics. The report would be accepted when the scores on all subabilities reached the level of adequate.

When scoring the reports of that semester we observed that students were not able to adequately improve their reports only based on the scores on the rubrics. To guide them we needed to add specific comments to the scores. These comments were sometimes specific corrections, which needed to be made, and sometimes corrections related to the data analysis, which was not covered by the rubrics. We also observed that we had difficulties providing feedback to groups who were assigned an application experiment. In an application experiment the results of one method should be compared to results from a different, independent method and the comparison must take into account experimental uncertainties. This part was often missing. In rubric B (designing and carrying out an experiment to investigate a phenomenon) there is no subability that addresses either of these. Therefore, we concluded that this rubric is not well suited for an application experiment and that we should probably include other existing Rutgers rubrics, which address

TABLE II. In the table, the "X" shows the rubrics that are used for each type of project.

	Rubrics				
Project	B	C	D	E	G
Observational	X			X	X
Testing		X		X	X
Application			X	X	X

application experiment related abilities, data analysis abilities, and others.

In the spring semester of 2015, we introduced the Rutgers rubrics in their original form (this marks the beginning of the rubrics period). Among the rubrics we selected the ones for the following five abilities: (B) "to design and conduct an experiment to investigate a phenomenon," (C) "to design and conduct a testing experiment," (D) "to design and conduct an application experiment," (G) "to collect and analyze experimental data," and (F) "to communicate scientific ideas." Since the rubric for the ability to communicate scientific ideas was developed for a written laboratory report, not a web-based report, we revised it to develop a new one specifically for a web-based report. We labeled it (E), since the original Rutgers rubrics do not contain a rubric labeled (E). We discuss its development in Sec. III D.

At the beginning of their work on the projects, each team of students received three rubrics as shown in Table II. The initial intention was to provide feedback to the students only in terms of rubric scores. We found later that specific comments were still necessary occasionally and in small numbers only for shortcomings that were so specific that they were not covered by the rubrics and it was unlikely for students to notice them only based on the scores on the rubrics (e.g., severe grammatical errors, or small notation lapses).

To make sure that the rubrics indeed covered all our main goals, we compared the statements in the adequate column of the rubrics with the elements that we paid attention to during the prerubrics period. Table III shows that all the listed elements are covered in the rubrics. The "correct physics" element is not mapped to a specific subability, because it is covered by more subabilities. Since the rubrics have been developed through research, this suggests that there is not a single overarching correct physics subability, but rather that physics is used in different ways in different subabilities. This emphasizes the fact that knowing correct physics is often context dependent, as many research have shown (see, for example, Ref. [40]). Also, some other rubric items (e.g., G4) do not match exactly with the prerubrics criteria. One reason is that the rubrics have not been developed by us, but adopted from Rutgers University and hence differences in the grouping or formulation of criteria are to be expected. Another important reason is that there has to be a balance between providing good and concrete guidance and the number of rubric items. Separating every single element of what it means to, for

TABLE III. A comparison between the subabilities in the prerubrics period (first column) and the subabilities in the rubrics period. Note that all criteria can be mapped to statements in the adequate column of the new rubrics.

PRERUBRICS criterion	SUBABILITY in the rubrics	ADEQUATE level
Clear description of the procedures. Use of figures	F1 Are able to communicate the details of an experimental procedure clearly and completely.	Diagrams and/or experimental procedure are clear and complete. Figures are appropriately chosen and correctly labeled. It takes no effort to comprehend.
Adequate labels on figures.	G4 Are able to record and represent data in a meaningful way	All important data are present, organized, and recorded clearly. The tables and graphs are correctly labeled and placed in a logical order.
Supporting claims	C8 Are able to make a reasonable judgment about the hypothesis.	A judgment is made, consistent with the experimental outcome, and assumptions are taken into account.
Correct data analysis (incl. expt. uncertainties)	G5 Are able to analyze data appropriately.	The analysis is appropriate, complete, and correct.
Correct physics	B9 Are able to devise an explanation for an observed pattern.	A reasonable explanation is made. It is testable and it explains the observed pattern.
	C4 Are able to make a reasonable prediction based on a hypothesis.	A prediction is made that follows from hypothesis, is distinct from the hypothesis, accurately describes the expected outcome of the designed experiment, and incorporates relevant assumptions if needed.
	D7 Are able to choose a productive mathematical procedure for solving the experimental problem.	Mathematical procedure is fully consistent with the design. All quantities are calculated correctly with proper units. Final answer is meaningful.

example, "represent data in a meaningful way" would be no different than providing concrete instructions on what to do. Instead, the rubrics are designed to be somewhat general so that students are forced to reflect on what it is that they did wrong. The reflection is the positive effect of rubrics on learning practices, which is mentioned the most in literature [28,33].

B. Sample and data collection

Our data consist of feedback given to each report in each iteration. The analysis was done on 61 reports, involving 297 students between the years 2008 and 2018. For 16 of the older reports we could not retrieve all the feedback, so we made approximations described in Sec. III C 1 b to include as many reports as possible. We only used reports assessed by one grader (S. F.) who remained the same throughout the considered period 2008–2018. The grader was trained by E. Etkina (E. E.) to use the rubrics in 2015. The training involves a novice and an experienced grader grading the same report and comparing and discussing their scores. In our case, the novice was S. F. and the experienced grader was E. E. She also helped establish reliability using 15% of the reports. The interrater agreement before discussion was about 80% and after discussion 100%.

The total number of reports that we could retrieve was 69. Out of these, 61 were included in the analysis. We excluded from the analysis all groups that have not completed the course (five in the prerubrics, none in the intermediate, and one in the rubrics period). One report was excluded because it contained more than one type of experiment and had to be evaluated on all the five rubrics instead of just three, which makes it an exception. Another report was excluded because the notes on the feedback and the content of the feedback could not be clearly understood. The remaining 61 reports were included in the analysis.

C. Procedure

1. Grader workload

a.The grading process.—We can conceptualize grading of a report to consist of three phases: (1) reading the report, (2) pondering the quality and contemplating the comments, and (3) writing the feedback. The times related to each phase will be called *reading time* (R), *pondering time* (P), and *writing time* (W), respectively. Their sum will be called the *total grading time* $T = R + P + W$.

The reading time is assumed to be the same in prerubrics and in the rubrics period. The writing time can be objectively assessed by the number of words in the feedback.

The pondering time is the most difficult to assess, especially since we do not have any objective data on this from the prerubrics period. Instead, we conducted a small investigation to assess the total grading time using two reports and four graders. The details are explained in Sec. III C 1 c.

Our goal was to determine the ratio between the grading times in the prerubrics period (index *comments*) and in the rubrics period (index *rubrics*). We were able to determine the ratios $W_{\text{rubrics}}/W_{\text{comments}}$ and $T_{\text{rubrics}}/T_{\text{comments}}$.

b.Number of words in a feedback.—The number of words in the feedback to one report was determined by counting the number of words in the feedback of all iterations for that report. This could be done on 17 reports from the prerubrics period, 6 reports from the intermediate period, and 21 reports from the rubrics period.

We had 16 additional reports from the prerubrics period available, but for these not all feedback was available. For 11 of these reports only the first feedback and the number of iterations was available, and for 5 of these reports only the second feedback and the number of iterations was available. In order to include these reports in the analysis we estimated the total word count for these reports in the following way. On the vertical axis of a graph we plotted the total word count for the 17 prerubrics reports for which the total word count was available. On the horizontal axis of the graph we plotted the number of iterations ($N_{\text{iterations}}$) times words in the first feedback (w_{first}) for the same 17 reports (Fig. 1). The plot shows an approximately linear relation with the slope 0.55. Using the function $0.55 N_{\text{iterations}} w_{\text{first}}$ (or $0.55 N_{\text{iterations}} w_{\text{second}}$ for 5 reports) we estimated the total number of words in the feedback for the 16 reports, for which we did not have all the feedback available.

c.Total grading time.—The writing time is an important part of the feedback process, but it does not reflect directly the total grading time. To quantify the total grading time with each method we did the following: (i) S. F. estimated the time spent grading using both methods. (ii) We asked S. F., Breena (a grader with a few years experience), Carl and Dennis (both novice graders) to grade two reports (I and II), one with the rubrics and the other with comments. The reports I and II were selected at random among the reports from the current year. No consideration was given to

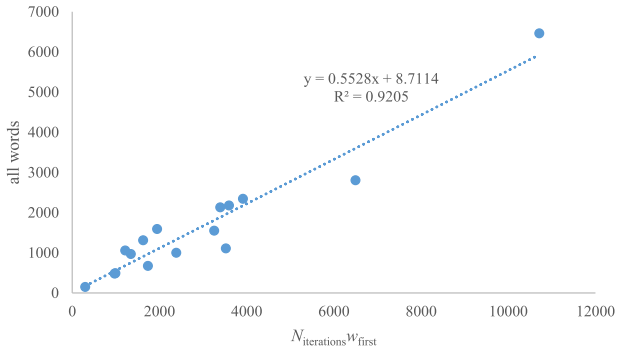

FIG. 1. The plot of total words versus words in the first feedback times the number of iterations. The coefficient in the fit was used to estimate the total words in the feedback for those reports, for which we could not retrieve all the feedback.

whether the reports are comparable. Table X in the findings section shows the graders, the graded reports, the grading methods, and the total grading times. These data allowed us to estimate the total grading time and calculate the ratio between rubrics total grading time and comments total grading time ($T_{\text{rubrics}}/T_{\text{comments}}$) for all combinations of graders and reports.

The collected data allowed us also to make a rough estimate of the ratio between the total grading time and the words in the feedback. We counted the words in the comments style feedback and calculated the ratio between the reported total grading time and the number of words in the feedback for all graders.

S. F.'s estimate could be potentially biased. We thus employed two methods to check the plausability of his estimate. The collected data allowed us to make a rough estimate of the ratio between the total grading time and the words in the feedback (T/w). We counted the words in the comments style feedback and calculated the ratio T/w for S. F. alone and the average of this ratio for all four graders and used each of them to estimate the total grading time based on the average number of words per iteration in the prerubrics period. The average number of words per iteration was calculated as $\sum w_{\text{report}}/(\sum N_{\text{iterations}} - N_{\text{reports}})$, where $\sum w_{\text{report}}$ is the number of all words in all iterations, $\sum N_{\text{iterations}}$ is the number of all iterations, and N_{reports} is the number of reports. The subtraction of the number of reports from the denominator is due to the fact that there is no feedback after the last iteration, just an acknowledgment of acceptance.

2. Quality of the reports

In the prerubrics period, it was the grader's professional decision whether a report was graded pass or fail. In the rubrics period an adequate score is ultimately required on all the rubrics. Note that all groups of students have opportunity for multiple revisions and resubmissions of their reports without penalties.

To assess the quality of the reports in the rubrics period (new reports) compared to the quality of the reports in the prerubrics period (old reports), we selected a random sample of six old reports that received a pass grade and evaluated them using rubrics. The pass grade means that the reports were considered adequate in the prerubrics period and we wanted to see if they would receive an adequate score also when graded using rubrics. An unweighted average was calculated from the numerical values of the scores received on each subability. Weighing was unnecessary, because any average below 3 would mean that the report would not have received a pass grade in the rubrics period.

3. Developing scientific abilities

We used the opportunity of scoring the old reports using rubrics to calculate the average score given to the old reports on each rubric item. This is a measure of which subabilities remained poorly developed in the prerubrics period.

To estimate which abilities were the most difficult for students to develop in the rubrics period, we analyzed the feedbacks of all iterations for each report and followed the progression of the score on each rubric item through the subsequent iterations. As a measure, we took the number of iterations after which the item received a full score. If the report received the full score on an item when it was first submitted, the value would be 1, not 0.

Based on previous research that indicates that traditional laboratories do not develop all of the scientific abilities assessed by the rubrics [15], we hypothesized that many of the difficulties that we might identify could be potentially explained by students being exposed to only cookbook, confirmational experiments. To substantiate this hypothesis at least a little, and in an attempt to minimize confirmation bias, we decided to approach this as a testing experiment. We gave the rubrics to an instructor who was not involved in the data analysis. The instructor was from our faculty and had over ten years of experience with traditional, mostly cookbook, confirmational university level physics laboratories for nonphysics majors. We asked him to predict which of the subabilities listed in the rubrics he thinks the students would have most problems with, based on the assumption that they have only been exposed to cookbook, confirmational experiments. We asked him to assign one of three difficulty levels (L–low, M–medium, H–high) to each subability. After receiving the scores, we wanted to account for the fact that we are scoring the report, not the actual process, therefore some subabilities will necessarily develop, if the project is to be completed. For example, the instructor will intervene if the students are unable to design an adequate experiment. Therefore, subability B2 (designing a reliable experiment) will necessarily receive an adequate score at first iteration and will, therefore, be considered (L). We assigned to each such subability the value (L), even if they have already been assigned some other value by the instructor. This was then the final prediction. The process is admittedly subjective, but it provides at least some form of testing our hypothesis.

D. The design of a new rubric

One of the nontraditional elements of the project laboratory course is the fact that our students submit a report in the form of a web page [41]. A web page based report is different from a written report in several ways. First, the report can be nonlinear. Hyperlinks enable branching. This also makes the report inherently difficult to print, hopefully saving some paper and thus reducing its impact on the environment. Second, the report is uploaded to a server and publicly available, which means that copyright laws must be taken into consideration. And third, the report can use a lot of web functionality and

TABLE IV. The first line of the new rubric for the ability of writing a report in the form of a web page.

Subability	0: Missing	1: Inadequate	2: Needs improvement	3: Adequate
E1 Are able to structure the web report in a way that it shows the important steps of the project	The web report does not load, is missing or is extremely unclear.	The web report is written as a monolithic story. It is difficult to access information about the different steps of the project.	The structure of the web report allows quick access to the different steps of the project, however, some important steps are missing or difficult to access.	The structure of the web report clearly shows all the important steps of the project and allows quick access to them.

the advantage of dynamic elements. As a consequence, we needed to develop a special rubric for the web report (rubric E). We started from the Rutgers rubric F for communication. The two items of the rubric deal with (F1) a clear and complete report and (F2) the evaluation of the purpose of the experiment and the meaning of the findings. These were merged to make item E3 on the new rubric. In collaboration with E. E. from Rutgers, we formulated the other items of rubric E for the ability to compose a web-based report about the findings (see Tables IV–VIII). This rubric has been revised during the course of its development and is given here in its form as of 2018 during the writing of this article.

Subability E1 (Table IV) concerns the nonlinear structure of the report. We expect all the important steps of the experiment to be available quickly, probably directly from the main menu.

Subability E2 (Table V) concerns the structure of the report. The usual introduction, methods, results, and discussion (IMRAD) structure common for traditional laboratory reports is not appropriate for all types of experiments. This line of the rubrics encourages students to use different report structures for different types of experiment. It is important for the students to understand different relationships between experiments, hypotheses, models, predictions, and evaluation in different types of experiments, and different flow of the thought process. In observational experiments one first observes, then measures, then produces models and then compares the model with the measurements (experiment, hypothesis, evaluation). In application experiments, the order is reversed. One first lists the knowledge and models on which the experiments are based. Then, one applies these to the problem at hand, and in the end compares the obtained results with the results from an independent source (model, experiment, evaluation). In the case of a testing experiment, one starts with the phenomenon and the hypotheses. Then, testing experiments are suggested and the predictions for the outcomes of these experiments are given, based on the various hypotheses. In the end, the experiments are performed and a judgment is made about the hypotheses based on the comparison between predictions

TABLE V. The second line of the new rubric for the ability of writing a report in the form of a web page.

Subability	0: Missing	1: Inadequate	2: Needs improvement	3: Adequate
E2 Are able to design the web report so that the type of the experiment (observational, testing or applicative) is clear from the structure and the order of the tabs in the navigation menu.	It is not possible to infer the type of the project from the structure of the web report.	The type of the project is listed in the web report, but its structure is not consistent with it (see adequate).	The structure of the web report is consistent with the type of the project, but some of the tabs should be reordered to better follow the structure of the type of the project. (See adequate.)	The structure of the web report and the order of the tabs are consistent with the type of the project. E.g., 1. In the case of an observational project, the tabs with the description of patterns, their mathematical representations, the explanations and the models come AFTER the tabs with the descriptions of the experiments. 2. In the case of a testing project, the tabs with the descriptions of the experiments and the predictions of the outcomes based on the tested hypotheses come BEFORE the description of the actual outcomes of the testing experiments. 3. In the case of an applicative project, the tab with the relevant theory and models comes first, THEN the tab with the description of the experiments and THEN the one with the analysis of data.

TABLE VI. The third and fourth lines of the new rubric for the ability of writing a report in the form of a web page. The subabilities E3a and E3b were joined together as E3 for some time during the rubrics period, but were soon split again into two separate subabilities.

	Subability	0: Missing	1: Inadequate	2: Needs improvement	3: Adequate
E3a	Are able to communicate the details of an experimental procedure clearly and completely.	The description and/or the pictures or sketches of the experimental setup are missing or unintelligible.	The descriptions of the experiments and procedures are unclear to the point where it is necessary to guess about some parts, perhaps because the pictures or sketches are missing or the structure of the text is such that it is difficult to follow the line of thought. The pictures or sketches are unclear or inadequately chosen (maybe some crucial ones are missing). The physical considerations contain major mistakes. Comprehending the descriptions takes a lot of effort.	The descriptions of the experiments, physical considerations, procedures and findings contain minor shortcomings. Pictures or sketches are included and adequately chosen, but contain minor shortcomings. Comprehending the descriptions takes some effort.	The descriptions of the experiments, physical considerations, procedures and findings are clear, complete and correct. The pictures or sketches are adequately chosen and labeled. The various representations are consistent with each other.
E3b	Are able to communicate the point of the experiment clearly and completely.	No discussion of the point of the experiment is present.	The point of the experiment is discussed, but vaguely. There is no reflection on the quality and importance of the findings.	The point of the experiment is presented clearly. Discussion of the quality and importance of the findings are superficial.	The point of the experiment is presented clearly. There is deep reflection on the quality and importance of the findings.

and outcomes (hypotheses–suggestion of experiment–prediction–experiment–evaluation).

In the original Rutgers rubrics, subabilities E3a and E3b (Table VI) are separate. In the development of rubric E we first joined them together into a single subability E3. However, we noticed that many students failed to adequately address the purpose of the experiment and the validity of the conclusions (E3b), but did well in describing everything in the experiment (E3a). A low score on the joined rubric E3 did not give them useful feedback about what they needed to develop more. Therefore, we separated again the two subabilities into E3a and E3b.

Subability E4 (Table VII) deals with the proper use of internet technology. Since the introduction of web reports in 2002, internet technology has changed multiple times in different ways. Throughout all these changes it became clear that insisting on the most basic web technology (pure HTML, no plug-ins), while allowing the spectrum of most important benefits (hyperlinks, thumbnails, colors, and menus) was the best method to make the reports available across multiple platforms. Here we had in mind high school teachers, who were among our target audience, and in our opinion could not be expected to always have access to up-to-date computers. So in this line of the rubrics we encourage hyperlinks to allow easier access to different parts of the report and to auxiliary data and resources. We also encourage the use of thumbnails (small, low resolution figures that link to higher resolution figures when clicked). On the other hand, we mention technologies that we find unproductive and unnecessary, such as fixed width text, the use of plug-ins (Flash, Java), and external scripts.

Subability E5 (Table VIII) addresses the copyright of pictures and other content from the Internet. We observed that it was tempting for students to just pick any useful picture from the Internet and use it in their report. Since the reports are published on the department webpage, we could not afford to have copyright infringement problems. Moreover, we found this to be a welcome opportunity to teach the students about the basics of copyright law. Therefore, we included copyright issues in the rubrics. We required the students to state under which license the material from the Internet was published and under which license they published their own work, especially pictures.

The rubric does not contain an item for the grammatical correctness of the language. Item E3a already covers clear communication, which includes adequate language.

The rubric evolved with use. We added examples of general, frequently observed shortcomings to the descriptions in the columns inadequate and needs improvement (see these columns in Table VII). We made similar slight modifications also to all the other rubrics, such that some, if now translated from Slovene back to English, might reflect a slightly different meaning or emphasis than the original Rutgers rubrics from which they were adapted. This reflects the importance of adaptation rather than just translation when using materials developed in different contexts.

TABLE VII. The fifth line of the new rubric for the ability of writing a report in the form of a web page.

	Subability	0: Missing	1: Inadequate	2: Needs improvement	3: Adequate
E4	Are able to use web technology appropriately.	The advantages of web technology are not used. If the report was printed out, it would loose nothing.	There is minimal use of web technology. The quality of the images is not adapted to viewing on screen (figures unnecessarily use a lot of memory or their resolution is too poor). If the report was printed out, it would loose only the benefit of cross hyperlinks. Unproductive or distractive use of web technology: e.g., excessive use of animated tabs and frames, use of Flash or other non-HTML elements for content that can be equally well presented using only HTML, difficulties in saving specific pages, unnecessary visual effects, parts of text or images overlap, etc. The page relies on external scripts to function properly. The page is not displayed correctly on some systems.	Some quotations are not hyperlinked to their citation. The page does not display entirely correctly (overlapping text, missing accented letters, superscripts, subscripts, size of images...). No thought was given to the page displaying correctly on different systems or even different screen resolutions.	Quotations are hyperlinked to citations. All hyperlinks offer the possibility to open in a new tab or window. All displayed images have a size below 200 kB. If necessary, they serve as thumbnails to larger versions. The authors gave some thought to how the page would display on tablets, mobile phones, and other systems. All scripts used by the page are stored locally. There are no automatic redirections.

TABLE VIII. The sixth and last line of the new rubric for the ability of writing a report in the form of a web page.

	Subability	0: Missing	1: Inadequate	2: Needs improvement	3: Adequate
E5	Are aware of copyright legislation.	There is no citation of sources of nonoriginal elements.	Most of the sources of nonoriginal material are cited, but among those without cited sources it is unclear whether they are original.	The sources of all nonoriginal material are cited. It is clear which source refers to which material. The sources of all material taken from the web are cited, but it is unclear under which license they were published. The original material does not contain licensing information.	The sources of all nonoriginal material are cited. It is clear which source refers to which material. The sources of all material taken from the web are cited, and it is clear under which license they were published on the web. The authors of the original material are stated and the information about the permissions for further use (licensing) is given.

E. Students' reception of the rubrics

Positive attitude is beneficial for learning. In the years 2015 and 2016 we collected students' reflections on their own learning after their project has been successfully competed. We did not specifically ask for their perception of rubrics, but some included it in their reflection. We analyzed these reflections and in the findings in Sec. IV D we present some of the students' quotes.

IV. FINDINGS

A. Grader workload

1. Number of words in the feedback

The number of words per group is shown in Fig. 2. With the introduction of the Rutgers rubrics in 2015 the total word count in the feedback changed from an average of 1549 to an average of 728 (0.69 standard deviations).

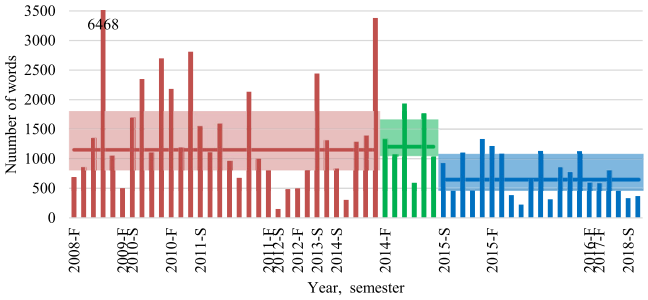

FIG. 2. An estimate for the number of words used to write all the feedback for each group. In the color graphic, red indicates the prerubrics period (before 2014-F), green indicates the intermediate period (the semester 2014-F), and blue indicates the rubrics period (from 2015-S on). Fifty percent of all reports fall within the respective shaded areas (the interquartile ranges). The lines indicate the respective medians. The labels on the horizontal axis indicate the year and semester [spring (S) and fall (F)].

The ratio is $W_{\text{rubrics}}/W_{\text{comments}} = 0.47$ and is statistically significant (t-test, two-tailed, $p < 0.01$). The median number of words changed from 1267 to 658 words per group. The ratio is $W_{\text{rubrics}}/W_{\text{comments}} = 0.52$. The Kruskal-Wallis test on the two distributions found that the two distributions are significantly different ($p < 0.01$). In the intermediate period, the average number of words was 1292 and the median was 1206. We use medians in addition to averages, because they are less susceptible to outliers.

2. Total grading time

To assess the total grading time we used two methods:

(i) S. F.'s total grading time in the rubrics period on a sample of six reports was $T_{\text{rubrics}} = 48 \pm 8$ min. He estimated his average grading time with comments in the prerubrics period as $T_{\text{comments}} = 90 \pm 15$ min. Therefore, $T_{\text{rubrics}}/T_{\text{comments}} = 0.53 \pm 0.24$. This estimate could be unintentionally biased. Table IX serves to test the plausibility of S. F.'s estimate and shows the total grading times (T) and the number of words in the comments of each grader (w). The ratio T/w for S. F. in the single measured report is 0.11 min/word. His average number of words per iteration in the prerubrics period was 570. This gives an estimated time of 63 min per iteration. Taking the average of all graders $T/w = 0.23 \pm 0.13$ min/word, one gets an estimate of 130 ± 70 min. Based on these results we find the estimate of 90 ± 15 min plausible.

(ii) All four graders graded two reports using either rubrics or comments. The total grading times, including the grader, the graded report and the grading method are shown in Table X. The average of all the ratios $T_{\text{rubrics}}/T_{\text{comments}}$ for all combinations of graders and reports is $T_{\text{rubrics}}/T_{\text{comments}} = 0.79 \pm 0.30$. This average has been compared to 1 with a t test and the difference is significant ($p < 0.01$). The Kruskal-Wallis test gives a $p = 0.19$. Excluding the data from S. F. due to potential bias, the avearage of the ratios is $T_{\text{rubrics}}/T_{\text{comments}} = 0.64 \pm 0.18$. A t-test comparison with 1 and a Kruskal-Wallis comparison of the distributions both give a $p < 0.001$.

TABLE IX. The table shows the grader, their reported total grading time, the number of words in their feedback, and the ratio T/w along with the average ratio T/w.

Grader	Time (min)	Words	T/w (min/word)
S. F.	40	378	0.11
Breena	90	209	0.43
Carl	60	285	0.21
Dennis	74	414	0.18
Average			0.23 ± 0.14

TABLE X. The table shows the grader, the report, the grading method and the total grading times. The method is indicated by font: roman font for rubrics and italic font for comments.

Grader	Report I	Report II
S. F.	*40 min*	50 min
Breena	40 min	*90 min*
Carl	*60 min*	40 min
Dennis	60 min	*74 min*

B. Quality of the reports

The average score that the sample of the old reports received on the new rubrics is shown in Fig. 3. Only one report received adequate on almost all rubric items (a numerical score of almost 3). The report 2011-S was scored twice. Once with the rubric for an observational experiment [2011-S(a)] and once with the rubric for a testing experiment [2011-S(b)]. The reason is that in the prerubrics period practically no project was specifically designed to be a testing experiment. This one came the closest and we wanted to see whether using a more suitable rubric would change the score. It improved it, but only slightly.

C. Estimating which subabilities are most difficult to develop

The distribution of scores per rubric item of the prerubric period reports are shown in Fig. 4 as a box-and-whiskers type diagram. All reports received an adequate score on several rubric items. However, a median of 0 was found for items B6 and D8 and a median of 1 for items G1, G2, and G3. Table XI lists all the items with a median of 2 or below.

In the rubrics period, a measure of the difficulty to develop a subability is the number of iterations required for a report to receive an adequate score on that rubric item, shown in Fig. 5. The closer the number is to 1 (adequate score at first hand-in), the easier the ability is to develop or the more it has been emphasized during previous instruction. The box-and-whiskers representation offers a natural set of criteria for sorting the scores into four discrete *difficulty levels*, based on a single threshold (see Fig. 5). We named the levels low difficulty (L), medium difficulty

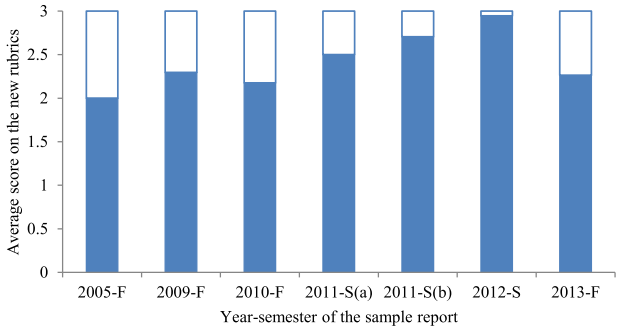

FIG. 3. Accepted old reports scored with the new rubrics. The figure shows the average score over all subabilities of each report.

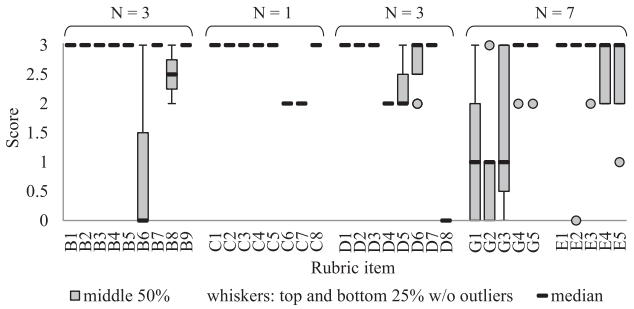

FIG. 4. The scores given to the old reports on each rubric item shown in a box-and-whiskers type diagram. The black line represents the median, it indicates the value under which at least 50% of all reports scored. The gray area is the interquartile range with at least 50% of the scores, and the whiskers are the remaining at most 50% of the scores. The outliers are represented as separate points (gray circles).[1] Rubric C was only used for one report, so there is no distribution of scores. Examples: On rubric item B6, the majority of the scores were 0 (at least 50%—median), while scores reached up to 3. Rubric items G2 and G3 have both at least 50% of the scores between 0 and 1, but G2 has only one outlier with a score of 3, while G3 has between 25% and 50% of the scores between 1 and 3. Therefore, although the median is the same, the distribution of G2 is peaked towards lower scores, while of G3 towards higher scores. Both, however, contain at least one score 3.

(M), high difficulty (H), and very high difficulty (V). The threshold was chosen such that we obtained a reasonable distribution between the values (L),(M), (H), and (V). For the threshold we chose the value 2 on the vertical axis (meaning adequate score at second iteration) and the criteria are as follows: (L) is assigned to all subabilities for which the gray area is entirely below the threshold (inclusive), meaning that 75% of reports achieved an adequate score by the second iteration. (M) is assigned to all subabilities for which the threshold is within the gray area, with the median equal or below it, meaning that 50% of reports achieved an adequate score by the second iteration. (H) is assigned to all subabilities for which the threshold is within the gray area, with the median above it, meaning that 50% of reports needed more than two iterations to achieve an adequate score. And (V) is assigned to all subabilities for which the gray area is entirely above the threshold (exclusive), meaning that 75% of reports needed more than two iterations to achieve an adequate score.

The abilities that appear difficult to develop, levels (H) and (V) are listed in Table XII.

[1]In a standard box-and-whiskers diagram, the *lower quartile* is the value below which 25% of all points lie and the *upper quartile* is the value below which 75% of all points lie. The *interquartile range* is the difference between the upper and lower quartile. The outliers are defined as points lying above the upper quartile plus 1.5 times the interquartile range and below the lower quartile minus 1.5 times the interquartile range. The "whiskers" extend to the lowest and highest point excluding outliers.

TABLE XI. The subabilities that received low scores on the reports of the prerubrics period. Only subabilities with a median below 2 are included.

Subability (in order of increasing median)
D8: Are able to determine how their assumptions affect the results obtained from the mathematical model.
B6: Are able to identify the shortcomings of the experiment and suggest improvements.
G2: Are able to estimate how particular experimental uncertainties affect the end result.
G1: Are able to recognize the sources of experimental uncertainties.
G3: Are able to describe how to minimize experimental uncertainties and they do it.
C6: Are able to identify how specific assumptions affect the result.
C7: Are able to decide whether the outcome agrees with the prediction.
D4: Are able to make a judgment about the outcome of an experiment or measurement.
D5: Are able to design a new, independent experiment to evaluate the results of the previous one.

The results of our prediction of difficulty levels are shown in Fig. 6. Three subabilities, B1, B2, and B3, were *post hoc* assigned an (L) value in the prediction, because the instructor would have intervened, if the students had not achieved an adequate level on these subabilities before the end of the practical part of the project. These are indicated by an "O" (for "overridden") in Fig. 6. The predictions are compared to the measured data from Fig. 5, where we replaced the measured values with the values (H), (M), and (L) as described above. Values (H) and (V) were grouped together. Figure 6 shows that 53% of the measured values match with the prediction. Among the rest, 22% had a lower measured difficulty level than predicted and 25% a higher measured difficulty level than predicted. The correlation coefficient between the predicted and measured

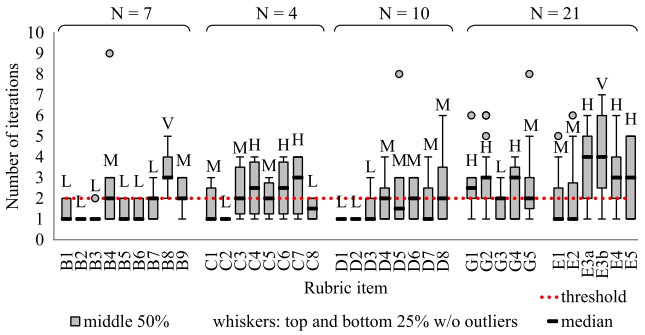

FIG. 5. The number of iterations needed before a report in the rubrics period received a full score on a particular rubric item represented as a box-and-whiskers type diagram. The difficulty levels are determined as described in the text. The threshold is indicated by the dotted line.

TABLE XII. The subabilities that take the most iterations to develop. Included are only subabilities assigned a difficulty level (d.l.) of H or V. See text for explanation.

Subability		d.l.
Observational experiment		
B8	Are able to represent a pattern mathematically (if applicable).	V
Testing experiment		
C4	Are able to make a reasonable prediction based on a hypothesis.	H
C6	Are able to determine specifically the way in which assumptions might affect the prediction.	H
C7	Are able to decide whether the prediction and the outcome agree or disagree.	H
Data analysis		
G1	Are able to identify sources of experimental uncertainty.	H
G2	Are able to evaluate specifically how identified experimental uncertainties may affect the data.	H
G4	Are able to record and represent data in a meaningful way.	H
Web report		
E3	Are able to communicate the details of an experimental procedure clearly and completely. Are able to reflect on the point of the experiment and the importance of the findings.	V
E4	Are able to adequately use web technology.	H
E5	Are aware of copyright laws.	H

values is 0.44. Among the subabilities that we predicted would have a high difficulty level, 44% were actually measured with a (H) or (V) difficulty level.

D. Students' reception of the rubrics

The students' perception of rubrics is reflected in the following quotes:

> *"... The rubrics were also useful, because they contain all the important elements of experimental work such as assumptions and experimental uncertainties…"*
> *"... When writing the report, we had to consider all the experiments again and all the details that could affect the results,… I learned a lot while thinking about all the parameters, such as control of variables and experimental uncertainties, and even more about considering them in general. In doing this the rubrics proved to be very useful…*
> *"... I think the basic idea of the course is very good, also the rubrics were very carefully designed and proved to be very helpful in our work..."*

One student who participated in the project laboratory course in both the prerubrics and the rubrics periods commented that he missed the detailed feedback of the prerubrics period.

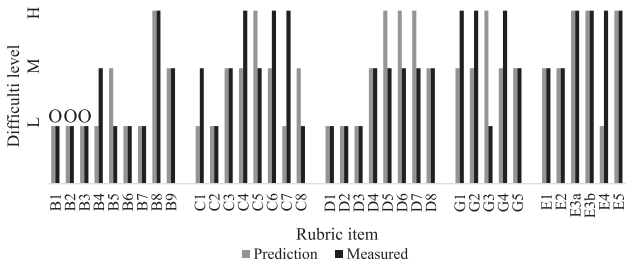

FIG. 6. Comparison of our predictions and the measured difficulties of students. The difficulty levels are as described in the text. The labels O indicate the overridden values, as described in the text.

> *"The first time that I was in the project laboratory course, we received 4 pages of comments, which encouraged some new ways of looking at the problem, while this time we only received numbers and a few characters in front of them. Which is not necessarily a bad thing, because it forces you to think about where did you go wrong and what needs to be fixed and with this offers you a new dimension. But, it can be time consuming..."*

These quotes are representative of the sample. Of particular importance is the last quote comparing the prerubrics period and the rubrics period feedback.

V. DISCUSSION

A. Grader workload

Earlier we conceptually divided grading a report into three phases. The first phase, reading the report, is assumed to take the same amount of time using either the comments method or the rubrics method.

The third phase, writing the feedback, was approximated with the number of words in the feedback. Figure 2 shows that the introduction of the rubrics greatly reduced the writing time of the grader. The overall ratio $W_{\text{rubrics}}/W_{\text{comments}}$ is approximately 50% and the change is statistically significant (t test, single tailed, $p < 0.01$; Kruskal-Wallis, $p < 0.01$).

Our analysis of the total grading time based on two reports and four graders (Table X) shows that the total grading time using the rubrics is on average $(79 \pm 30)\%$ of the total grading time using comments (t test, single tailed, $p < 0.01$), and is as low as $(64 \pm 18)\%$ when excluding S. F. due to potential bias (t test, single tailed, $p < 0.001$, and Kruskal-Wallis, $p < 0.001$.). All other methods of calculating either average ratio give similar results, which supports the robustness of both results.

The anecdotal result from S. F. in his 11-years experience indicate an even bigger difference ($T_{\text{rubrics}}/T_{\text{comments}} = 0.53$). While there might be some unintentional bias in

this result, the result is within 1 standard deviation of both results reported above. A larger scale investigation in the future could give more reliable results.

In Table X, there is one instance of the grading time with comments being shorter than the grading time with rubrics (S. F.). We have no particular explanation for this instance, but we note that it is not entirely unexpected. Similar instances can be observed also for the number of words in Fig. 2 (for example 2012-F and 2015-F). Despite this, the average ratio $T_{\text{rubrics}}/T_{\text{comments}}$ is still statistically significantly different from 1. Even more so, if the times of S. F. are excluded.

The most difficult time to assess is the pondering time. On one hand, using rubrics might make a novice grader often go back and forth between the rubrics and the report to decide what grade to give on a particular subability.

On the other hand, when not using rubrics, the commitment to consistency in grading, i.e., using the same criteria for grading every report, often makes the grader go back to the reports that they have already graded and correct those grades based on new data from later reports.

It is arguable how similar these times are and which is shorter, but since the total grading time is shorter using rubrics, any difference in the pondering time does not compensate for the difference in the writing time.

These findings show that using the rubrics can substantially reduce the total grading time. A more thorough investigation with a larger sample of reports and graders remains for future research.

We suppose that due to the decrease in the workload of the grader, there is probably an increase in the workload of the students, who have to first identify the concrete shortcomings before correcting them. A quote from a student presented in Sec. IV D seems to support this supposition.

B. Quality of the reports

From Fig. 3 we can conclude that the quality of the reports in the rubrics period increased. As mentioned in Sec. III D, the only feature of the reports that was possibly better in the prerubrics period is the grammatical correctness of the language. To test that grammatical comments did not contribute significantly to the prerubrics feedback length, we extracted the linguistic comments from seven randomly selected feedbacks from the prerubrics period (21% of the prerubrics period reports). The maximum percentage of words in a feedback that these comments covered was 2.3%, with four feedbacks containing no such comments. The average was 0.5% of all words in a feedback. Therefore, we conclude that the mere exclusion of grammatical comments from the rubrics period feedback did not significantly affect the observed decrease in the workload.

C. Developing scientific abilities

1. The subabilities difficult to develop

Figure 4 shows the scores that the accepted reports from the prerubrics period received when scored with the rubrics. The subabilities that received a full score have been obviously well developed even before the introduction of the rubrics. This could be due to the course work students have taken prior to the project laboratory course or due to the emphasis during the course itself. The subabilities that received low scores have been obviously poorly developed before the introduction of the rubrics. It appears that the grader did not consider them important during the scoring, or was perhaps even unaware of them. These subabilities are shown in Table XI. In addition to being a measure of which subabilities the instructors might have neglected, this is also an indicator of the subabilities that are likely among the more difficult to develop, even if this is simply due to students and/or instructors not paying much attention to them during instruction.

A better measure of the difficulty level of the subabilities is the number of iterations needed for a subability to be scored as adequate, which was measured in the rubrics period and is shown in Fig. 5. Some of the highest scoring (most difficult to develop) subabilities are shown in Table XII. A comparison between Table XI and XII, as well as Figs. 4 and 5 reveals that the following subabilities are difficult to develop according to both criteria: G1 (sources of experimental uncertainties), G2 (how uncertainties affect the result), C6 (how assumptions affect the prediction of a model), and C7 (deciding whether the prediction and outcome agree). Moreover, subabilities D4 (making a judgment about the outcome of an experiment or measurement) and D8 (how assumptions affect the result of a mathematical model) are borderline difficult according to the criterion in Fig. 5 (the median is at 2, but not above). This indicates that many of the subabilities, which have been poorly developed in the prerubrics period, appeared difficult to develop also in the rubrics period. On the other hand, subabilities B6 (identify shortcomings and suggest improvements) and G3 (minimizing experimental uncertainties), both among the poorly developed in the prerubrics period, have been quickly developed in the rubrics period. This suggests that these two subabilities are easily developed when students are provided with clear expectations. In the following paragraphs we discuss the possible reasons for the difficulties.

All subabilities in the G rubric, which have high or very high scores on the difficulty level scale, refer to experimental uncertainties. Subability G1 deals specifically with identifying the experimental uncertainties. Students, enrolled in the project laboratory course, have all had at least three years of physics in high school (210 hours in total) with at least 30 h of experimental activities in total. The estimation of and calculation with experimental uncertainties is included in the high school curriculum.

However, in school practice, the experimental uncertainties are rarely estimated from a single measurement and are most often of a statistical nature and calculated via repeated measurement. This could explain why identification of other types of uncertainties, such as systematic ones, and estimation of uncertainties from single measurements were found to be more difficult to develop. For example, Munier, Merle, and Brehelin [42] found that almost all elementary school students in their research considered commercial measurement devices free of systematic uncertainty. Similarly, Allie *et al.* [43] report that 51% of first year university students in their study did not take into account systematic uncertainty.

From the content analysis of some of the reports, we observed that much of the difficulties with G2, evaluating the uncertainty of the final result, arise from students not explaining adequately how they arrived at the reported uncertainty, rather than having it calculated wrongly. Some additional difficulties probably arise from the fact that most high school experiments are of the cookbook type. In such experiments, the experimental setup is predetermined and instructions often include the number of measurements to do. Therefore, students did not learn in high school how to make these kinds of decisions, because the decisions were made for them. Several studies indicate that the idea of a measurement result being a single value is very persistent and special educational interventions are required to address the role of experimental uncertainties [42–45]. Students can, however, still reduce the experimental uncertainties by carefully reading out the equipment output (avoiding parallax etc.) and carefully controlling initial (and all other) conditions.

The entire ability C, which on average scored high on the difficulty level scale in the rubrics period, deals with the testing experiments. This type of experiment is practically never present in high school, which would explain a higher difficulty level. The items C4 and C7 deal with the essence of a testing experiment: make predictions and evaluate the outcome with regard to the prediction. Neither of the two subabilities are typically addressed in high school physics courses. Tasquier, Levrini, and Dillon [46] report that in the middle of an intervention about the role of models in physics, only 1 out of 26 students mentioned that models are used to predict an outcome and 2 out of 26 students mentioned that "model is for testing, for example, 'Model is a prototype for testing a phenomenon'". The study by Arnold *et al.* [45] focuses specifically on testing experiments and among the tasks identified by the researchers, the task of predicting the outcome based on the hypothesis is not included. Item C6 deals with assumptions. The high difficulty level is likely due to students not being used to even thinking about them. A study by Slisko and Corona Cruz [47] found that high school students only thought about assumptions when their results were surprising and they could not find any other explanation. The study by Arnold *et al.* [45] does not address assumptions specifically, but it includes controlling for "confounding variables," which could be arguably thought of as variables that are assumed to remain constant or to be irrelevant for the outcome of the experiment. Arnold *et al.* found that 75% of students did not consider these variables at all. High school experiments and physics courses in general rarely emphasize the limitations of the models that are being taught. Therefore, students seldom experience the failure of a model due to its assumptions not being valid and hence overlook the role of the assumptions.

Subabilities D4 and D5 deal with the evaluation of an experiment. If students have always performed only confirmational experiments, they have never been exposed to a situation where they do not know the "correct" answer and have to figure it out in an independent way. The study by Allie *et al.* [43] report that at best 30% of the students in their sample are aware of the importance of experimental uncertainties when comparing two datasets. However, they report that only 15% use this knowledge consistently.

Our students have probably also never been exposed to a situation where the results of an experiment do not match the predictions and a rejection of a hypothesis or model is presented to them as a legitimate option. A study by Girault *et al.* [48] found that although in their sample only 33% of laboratory instructions could be considered completely cookbook, 81% of laboratories still only had one way of completing the task. So not only do ill-structured instructions still appear to be close ended, they also do not appear to address the possibility of a hypothesis being rejected. Therefore, it is understandable that these subabilities took longer to develop.

The ability E deals with the web report. The subability E3a includes the descriptions of everything. As long as there is anything remaining unclear in the descriptions, this item cannot receive a full score. If students have indeed mostly performed cookbook experiments prior to the project laboratory course, the descriptions would have already been included in the material they received. Therefore, the ability to describe their own experiment has typically not been an explicit learning goal thus far. Subability B8 is similar to E3a in the sense that a full score requires a correctly developed mathematical model. As long as there are any shortcomings in it, the item cannot receive a full score. Confirmational experiments do not help develop this subability, since the mathematical model is given in advance. One study shows that only 38% of students were able to create a mathematical model from a given situation without errors [49]. Other studies show that special activities are required to develop modeling skills [50–53].

The subability E3b requires a short reflection on the experiment: What did the students attempt to achieve with it and how well did they achieve it? Specifically, we want them to evaluate their conclusions. How solid are they? If the students have mostly done confirmational experiments

in the past, they have most likely never been asked to defend their findings since they already matched with the established theory. And, if the outcomes did not match the theory, the students were usually asked to explain why the experiment "went wrong," and not to defend their actual findings. This may explain why it is difficult for the students to evaluate the validity of their conclusions. In fact, many do it by comparing them with whatever they think is the theoretical result, instead of judging the merit of their method on its own.

The subability E4 refers mostly to the use of thumbnails, hyperlinks, avoiding unnecessary animated elements, like blinking menus, etc. Students have the most problems with the use of hyperlinks to link between pages. Rubric E encourages easy access to various parts of the report. We expect that every time a part of the report is referenced, it would also be hyperlinked, but students rarely do this.

Figure 6 indicates that approximately half of the difficulties that the students have may be expected based on the assumption that they have only been exposed to cookbook, confirmational experiments. This is consistent with the findings of Etkina *et al.* [15] that traditional laboratories do not develop all the sub-abilities assessed by the rubrics, and the study by Berg *et al.* [23], which shows that open-ended experiments significantly increase learning gains at higher taxonomy levels (application, analysis, synthesis, and evaluation) in comparison with traditional experiments.

For three subabilities in Fig. 6 our predictions and the measured values differ by two difficulty levels. These are summarized in Table XIII. Here we comment on them.

The subability C7 deals with matching the outcome of an experiment with the prediction of the outcome. It appeared to us that students should be able to do this easily once the prediction has been made. As it turned out that making predictions has a high difficulty level (see subability C4 in Fig. 5), it may be that predictions have been made in a way that was not helpful in matching the outcomes to them. Only after the predictions have been corrected, the matching could be done properly. In fact, within the four testing experiments in our sample, in three cases the subability C7 received the score of adequate in the same iteration as C4 and in one case one iteration later. This is consistent with the explanation above.

TABLE XIII. A list of abilities for which there is a two level difference between our prediction and the measured difficulty levels.

Subability		Pred.	Meas.
C7	Are able to decide whether the prediction and the outcome agree or disagree.	L	H
G3	Are able to describe how to minimize experimental uncertainty and actually do it.	H	L
E4	Are able to adequately use web technology.	L	H

The subability G3 deals with minimizing the experimental errors. We expected that this would be hard for students, since in cookbook experiments most parameters that determine the experimental uncertainties (the setup, the equipment to use and the number of measurements to do) are decided in advance. It appears that our population of students nonetheless payed adequate attention to these parameters.

We assigned E4 (technical aspects of the report) a low value, because we believed, and still do, that the description in this rubric item is so clear and specific that it should not be a problem following it. At the moment we have no explanation why this subability would have a high difficulty other than the possibility that the students simply left it for last.

2. Comparison with previous studies

We compare our findings about the difficulty level of the subabilities with those in the report by Etkina, Karelina, and Ruibal-Villasenor [16], because our rubrics were adapted from the rubrics used therein, making the comparison particularly meaningful. In the study by Etkina, Karelina, and Ruibal-Villasenor, the analysis of six subabilities is presented. Their setup is different from ours. They have studied multiple groups of students engaged in a series of 12 consecutive laboratory activities in one course. All groups engaged in the same laboratory at the same time and then moved on to the next laboratory. Each report for each laboratory activity was assessed only once. For comparison with our study, we had to assign the difficulty level of a specific subability. We based it on the number of laboratory activities it took before at least 80% of the reports received a score of needs improvement or adequate on that subability. The criterion was chosen such that it gave a reasonable distribution of (L),(M), and (H) values, and was the following: if 80% of the reports received the score of needs improvement or adequate by laboratory No. 4 (inclusive), we assigned (L), if by laboratory No. 7 (inclusive), we assigned (M), if later, we assigned (H). The assigned difficulty level values of the six subabilities are in Table XIV.

From Table XIV we can see that only one subability has a larger than 1 difference in difficulty level. There is a one level difference for subabilities D7 and D8, both dealing with assumptions. It is not surprising that identifying assumptions and their effects in a setup where the context changes every time is more difficult than in a setup where one repeatedly tries to identify them in one context. An unexpectedly large difference is on subability G1, which deals with identifying experimental uncertainties. Etkina, Karelina, and Ruibal-Villasenor [16] report that in their study there was special emphasis placed on the uncertainties in the first lab. This may explain the difference.

D. The effect of rubrics on the teachers and the students

We observed a positive effect of the rubrics on our instruction. Figure 3 shows that some reports, which have

TABLE XIV. A tentative comparison between the difficulty levels of some subabilities as reported by Etkina, Karelina, and Ruibal-Villasenor [16] and as determined in our study. We used three difficulty levels as explained in the text.

		Reference [16]	Our study
C3	Are able to differentiate between an explanation and a prediction.	M/H[a]	M
D5	Are able to evaluate the results by means of an independent method.	M	M
D7	Are able to identify the assumptions made in using the mathematical procedure.	M	L
D8	Are able to determine specifically the way in which assumptions might affect the results.	H	M
G1	Are able to identify sources of experimental uncertainty.	L	H
G2	Are able to evaluate specifically how identified experimental uncertainties may affect the data.	H	H

[a]The data is reported in Ref. [16] differently from the other subabilities, so the same criterion to assign L, M, or H as for the other subabilities could not be applied. Based on the data, two assignments are reasonably possible.

received a passing grade in the pre-rubrics period, would not meet the criteria for passing in the rubrics period. This indicates that in the pre-rubrics period even the instructors have not been sufficiently paying attention to some sub-abilities. If even we have not noticed these shortcomings in the reports, we have probably not payed any attention to them during the course and neglected to develop them.

The rubrics also made us explicitly aware of the importance of testing experiments. As mentioned above, in the pre-rubrics period, we have never designed an entirely testing project task. We considered it part of the observational experiment and never gave it special emphasis. Thus, students have likely never encountered a specifically testing experiment before, which explains why it was somewhat difficult for our students to develop the abilities associated with testing experiments. Only after focusing additionally on teaching the epistemology of physics did we realize the importance of the testing experiments and the ability to perform them as a separate and important ability. Most groups of students who were given a testing project wanted to start out by looking for known "theory" on the topic and applying it to evaluate the suggested hypotheses. This "theory-first" approach is not unexpected, since the usual process of exploring a scientific question is to first look for literature already published on the topic. Students often asked us, why should they perform or even design a testing experiment, if it is "clear" from the theory, which of the hypotheses is correct. Since these are physics students, the obvious answer is that they will need this ability in the future when they make new discoveries and need to test their innovative ideas, for which there might be no theory to consult. However, the question itself reveals that they are not aware that designing a testing experiment in itself is a relevant learning goal. They seem to focus on the specific content goals rather than the general ability being developed in the process of achieving the content goals. This might help explain the consistent findings here and in Ref. [16] that the ability to design a testing experiment is among the most difficult abilities to develop. Students do not appear to be aware that this is an ability in itself.

However, students like to see the usefulness of what they are doing in the short term, therefore, providing an immediate relevant context might be beneficial. One such context is related to the communication of science. When the general public are confronted with an idea that they might not find intuitive or have heard competing hypotheses about, which seem more intuitive to them, they might not be convinced by a scientist simply saying that: "It follows from all we already know." Therefore, designing a testing experiment that shows especially the invalidity of some of the competing hypotheses might be an important tool for the communication of science, and especially the communication of the epistemology of science to a general audience. This is an ability that is useful for any scientist to have.

The students' positive perception of the rubrics is an important element in making them an efficient tool for learning. The positive reception can be seen in the statements quoted in Sec. IV D.

The last comment mentions that they appreciated the longer comments of the prerubrics period, but find also value in the self-reflection required by using rubrics. This seems to be in agreement with the findings of Ene and Kosobucki [29] who reported that in an essay assignment the rubrics were valued, but concrete comments were valued more. They also mention that using the rubrics forces self-reflection, but it can be time consuming. This seems to confirm our supposition that due to less workload on the grader, there is somewhat more workload on the students.

VI. CONCLUSIONS

A. Answers to research questions

We have described the introduction and use of assessment and self-assessment scientific abilities rubrics in a project-based course. With the study, we answered the research questions as follows:

(i) To what extent do scientific abilities rubrics help to reduce the grader's workload in the project laboratory course?

The use of rubrics can reduce the writing time of the grader to approximately 50%. Evidence from a

short investigation shows that it can reduce the total grading time to approximately 75%. Anecdotal evidence from one grader indicates a possibly even greater decrease in total grading time.

(ii) How do scientific abilities rubrics affect the quality of students' work in open-ended project laboratory tasks?

The use of rubrics increases the quality of students' reports compared to the reports accepted when feedback was given in the form of comments.

(iii) How quickly do students achieve certain abilities and can the scientific abilities rubrics be used to evaluate this?

The abilities that take the most time to learn are handling experimental uncertainties and conducting a testing experiment. Among the other abilities, there are some subabilities that take longer to develop than others. These are mainly those related to conveying a clear description of the methods and procedures used, and those related to assumptions. We believe that the scientific abilities rubrics can be used to determine how quickly students achieve each ability by following the progression of the scores either in improving one lab report or through multiple subsequent lab reports.

In addition we found the following:

(iv) Providing students with the rubrics and the explanation of how they are used helps them quickly acquire subabilities related to the optimization of the experiment and the minimization of experimental uncertainties.

(ii) Rubrics are useful also for the instructors, because they clearly formulate learning goals of the course, which then guide the instruction. Specifically, without the rubrics, in the prerubrics period, we have been paying too little attention to some important aspects of scientific inquiry, such as designing testing experiments, the epistemological role of testing experiments, and the role of the assumptions in evaluating the results of both testing and application experiments.

We also propose that many of the students' difficulties in achieving adequate scores on the rubrics could be potentially explained by the fact that in all their prior education, our students mostly had experience with cookbook, confirmational experiments, which do not develop many of the subabilities evaluated with the rubrics. With this explanation we managed to predict 44% of the students' difficulties, assuming we correctly anticipated which subabilities such experiments would and would not develop.

B. Implications for instruction

Based on our conclusions, we recommend the use of rubrics, especially in project-based course work. Our experience with the adaptation and the development of new rubrics taught us that introducing rubrics as tools for assessment and self-assessment is an iterative process. We suggest starting out with validated rubrics and translating them, if necessary, and then gradually modifying them to best suit the needs of the course. The work of Nadji and Lach [54] and Buggé and Etkina [17] shows that this can be efficiently done also in high school. If more teachers teach the same or similar courses, we recommend that they develop the rubrics together. This can help clarify the learning goals of the course for the instructors and also help develop the best possible phrasing of the rubrics themselves. Students are different and might interpret the meaning of a statement in different ways, not necessarily as intended by the instructor. While such occurrences will eventually be corrected in the iterative process, having multiple developers can help correct them already during preparation.

ACKNOWLEDGMENTS

We would like to thank Bor Gregorčič for the valuable comments on the manuscript and Andreja Šarlah for help with the data analysis.

APPENDIX: THE SCIENTIFIC ABILITY RUBRICS

The Rutgers scientific abilities rubrics are freely available [11]. Each rubric represents a scientific ability.

A detailed description of the rubrics can be found in Ref. [13]. We give here one example. The rubric in Fig. 7 assesses the ability to design and conduct an observational experiment and is the source for our rubric to assess the ability to design and conduct an experiment to investigate a phenomenon. We have already mentioned that when transferring to our environment some modifications were necessary, so our rubric B is slightly different from the one in Fig. 7. We present here the source that is freely available in English in Ref. [11], and therefore easily accessible to larger audience.

Each row in the rubric represents a subability, which is an aspect of the ability. The columns describe the proficiency levels: 0—"missing, 1"—"inadequate," 2—"needs improvement," and 3—"adequate." In the first column of each row, there is the description of the subability. In the other columns, there are short descriptions of what would be considered evidence that the subability has been expressed at a certain level in the assessed work. These descriptions can provide guidance to students on how to improve their abilities. The most important is the last column, which represents the evidence that the ability has been adequately developed. The descriptions in the cells of this column are used as guidelines to describe what is expected of the students.

RUBRIC B: Ability to design & conduct an observational experiment					
	Scientific Ability	Missing	Inadequate	Needs improvement	Adequate
B1	**Is able to identify the phenomenon to be investigated.**	No phenomenon is mentioned.	The description of the phenomenon to be investigated is confusing, or it is not the phenomena of interest.	The description of the phenomenon is vague or incomplete.	The phenomenon to be investigated is clearly stated.
B2	**Is able to design a reliable experiment that investigates the phenomenon.**	The experiment does not investigate the phenomenon.	The experiment may not yield any interesting patterns.	Some important aspects of the phenomenon will not be observable.	The experiment might yield interesting patterns relevant to the investigation of the phenomenon.
B3	**Is able to decide what physical quantities are to be measured and identify independent and dependent variables.**	The physical quantities are irrelevant.	Only some of physical quantities are relevant.	The physical quantities are relevant. However, independent and dependent variables are not identified.	The physical quantities are relevant and independent and dependent variables are identified.
B4	**Is able to describe how to use available equipment to make measurements.**	At least one of the chosen measurements cannot be made with the available equipment.	All chosen measurements can be made, but no details are given about how it is done.	All chosen measurements can be made, but the details of how it is done are vague or incomplete.	All chosen measurements can be made and all details of how it is done are clearly provided.
B5	**Is able to describe what is observed without trying to explain, both in words and by means of a picture of the experimental setup.**	No description is mentioned.	A description is incomplete. No labeled sketch is present. Or, observations are adjusted to fit expectations.	A description is complete, but mixed up with explanations or pattern. The sketch is present but is difficult to understand.	Clearly describes what happens in the experiments both verbally and with a sketch. Provides other representations when necessary (tables and graphs).
B6	**Is able to identify the shortcomings in an experimental and suggest improvements.**	No attempt is made to identify any shortcomings of the experimental.	The shortcomings are described vaguely and no suggestions for improvements are made.	Not all aspects of the design are considered in terms of shortcomings or improvements.	All major shortcomings of the experiment are identified and reasonable suggestions for improvement are made.
B7	**Is able to identify a pattern in the data.**	No attempt is made to search for a pattern.	The pattern described is irrelevant or inconsistent with the data.	The pattern has minor errors or omissions. Terms proportional are used without clarity- is the proportionality linear, quadratic, etc.	The patterns represents the relevant trend in the data. When possible, the trend is described in words.
B8	**Is able to represent a pattern mathematically (if applicable).**	No attempt is made to represent a pattern mathematically.	The mathematical expression does not represent the trend.	No analysis of how well the expression agrees with the data is included, or some features of the pattern are missing.	The expression represents the trend completely and an analysis of how well it agrees with the data is included.
B9	**Is able to devise an explanation for an observed pattern.**	No attempt is made to explain the observed pattern.	An explanation is vague, not testable, or contradicts the pattern.	An explanation contradicts previous knowledge or the reasoning is flawed.	A reasonable explanation is made. It is testable and it explains the observed pattern.

FIG. 7. The rubric for the ability to design and conduct an observational experiment, which corresponds to our rubric for the ability to design and conduct an experiment to investigate a phenomenon.

[1] E. Etkina, Millikan award lecture: Students of physics–Listeners, observers, or collaborative participants in physics scientific practices?, Am. J. Phys. **83**, 669 (2015).

[2] N. G. Holmes, J. Olsen, J. L. Thomas, and C. E. Wieman, Value added or misattributed? a multi-institution study on the educational benefitof labs for reinforcing physics content, Phys. Rev. Phys. Educ. Res. **13**, 010129 (2017).

[3] F. Reif and M. St. John, Teaching physicists' thinking skills in the laboratory, Am. J. Phys. **47**, 950 (1979).

[4] M. F. Masters and T. T. Grove, Active learning in intermediate optics through concept building laboratories, Am. J. Phys. **78**, 485 (2010).

[5] A. Szott, Open-ended laboratory investigations in a high school physics course: The difficulties and rewards of implementing inquiry-based learning in a physics lab, Phys. Teach. **52**, 17 (2014).

[6] N. G. Holmes and C. E. Wieman, Introductory physics labs: We can do better, Phys. Today **71**, 38 (2018).

[7] A. Walker and H. Leary, A problem based learning meta analysis: Differences across problem types, implementation types, disciplines, and assessment levels, Interdiscip. J. Problem-Based Learn. **3**, 12 (2009).

[8] J. Strobel and A. van Barneveld, When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms, Interdiscip. J. Problem-Based Learn. **3**, 44 (2009).

[9] M. F. Di Mauro and M. Furman, Impact of an inquiry unit on grade 4 students' science learning, Int. J. Sci. Educ. **38**, 2239 (2016).

[10] Y. Doppelt, Implementation and assessment of project-based learning in a flexible environment, Int. J. Technol. Des. Educ. **13**, 255 (2003).

[11] Rutgers Physics and Astronomy Education Research group, Scientific abilities, acquired May, 22, 2016 https://sites.google.com/site/scientificabilities/.

[12] E. Etkina, A. VanHeuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[13] E. Etkina, D. T. Brookes, and G. Planinsic, *Investigative Science Learning Environment* (Morgan & Claypool Publishers, San Rafael, CA, USA, 2019), pp. 2053–2571.

[14] E. Etkina and A. VanHeuvelen, Investigative science learning environment—a science process approach to learning physics, *Research-Based Reform of University Physics* Vol. 1

(2007), https://www.compadre.org/Repository/document/ServeFile.cfm?ID=4988&DocID=239.
[15] E. Etkina, A. Karelina, M. Ruibal-Villasenor, R. Jordan, D. Rosengrant, and C. Hmelo-Silver, Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories, J. Learn. Sci. **19**, 54 (2010).
[16] E. Etkina, A. Karelina, and M. Ruibal-Villasenor, How long does it take? A study of student acquisition of scientific abilities, Phys. Rev. ST Phys. Educ. Res. **4**, 020108 (2008).
[17] D. Buggé and E. Etkina, Reading between the lines: lab reports help high school students develop abilities to identify and evaluate assumptions, in *Proceedings of the Physics Education Research Conference 2016, Sacramento, CA, 2016* (2016), https://www.compadre.org/Repository/document/ServeFile.cfm?ID=14191&DocID=4543.
[18] K. M. McGoldrick and B. Peterson, Using rubrics in economics, Int. Rev. Econ. Educ. **12**, 33 (2013).
[19] S. Allie, A. Buffler, L. Kaunda, and M. Inglis, Writing-intensive physics laboratory reports: Tasks and assessment., Phys. Teach. **35**, 399 (1997).
[20] C. Wieman and N. G. Holmes, Measuring the impact of an instructional laboratory on the learning of introductory-physics, Am. J. Phys. **83**, 972 (2015).
[21] L. E. Strubbe, J. Ives, N. G. Holmes, D. A. Bonn, and N. K. Sumah, Developing student attitudes in the first-year physics laboratory, Proceedings of the Physics Education Research Conference 2016, Sacramento, CA (2016), https://www.compadre.org/Repository/document/ServeFile.cfm?ID=14265&DocID=4619.
[22] B. R. Wilcox and H. J. Lewandowski, Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 020132 (2016).
[23] C. A. R. Berg, V. C. B. Bergendahl, B. Lundberg, and L. Tibell, Benefiting from an open-ended experiment? a comparison of attitudes to, and outcomes of, an expository versus an open-inquiry version of the same experiment, Int. J. Sci. Educ. **25**, 351 (2003).
[24] A. Mason and C. Singh, Do advanced physics students learn from their mistakes without explicit intervention?, Am. J. Phys. **78**, 760 (2010).
[25] H. G. Andrade, Student self-assessment: At the intersection of metacognition and authentic assessment, (1999), in *Paper Presented at the Annual Meeting of the American Educational Research Association* (Montreal, Quebec, Canada, April 19-23, 1999).
[26] E. Panadero and A. Jonsson, The use of scoring rubrics for formative assessment purposes revisited: A review, Educ. Res. Rev. **9**, 129 (2013).
[27] A. Alsina, S. Ayllón, J. Colomer, R. F.-P. na, J. Fullana, M. Pallisera, M. Pérez-Burriel, and L. Serra, Improving and evaluating reflective narratives: A rubric for higher education students, Teach. Teach. Educ. **63**, 148 (2017).
[28] A. Jonsson and G. Svingby, The use of scoring rubrics: Reliability, validity and educational consequences, Educ. Res. Rev. **2**, 130 (2007).
[29] E. Ene and V. Kosobucki, Rubrics and corrective feedback in ESL writing: A longitudinal case study of an L2 writer, Assess. Writ. **30**, 3 (2016).
[30] A. Cockett and C. Jackson, The use of assessment rubrics to enhance feedback in higher education: An integrative literature review, Nurs. Educ. Today **69**, 8 (2018).
[31] R. J. Howell, Exploring the impact of grading rubrics on academic performance: Findings from a quasi-experimental, pre-post evaluation, J. Excellence Coll. Teach. **22**, 31 (2011).
[32] W. D. Schafer, G. Swanson, N. Bené, and G. Newberry, Effects of teacher knowledge of rubrics on student achievement in four content areas, Appl. Meas. Educ. **14**, 151 (2001).
[33] P. M. Sadler and E. Good, The impact of self- and peer-grading on student learning, Educ. Assess. **11**, 1 (2006), https://www.tandfonline.com/doi/abs/10.1207/s15326977ea1101_1.
[34] A. Rajapaksha and A. S. Hirsch, Competency based teaching of college physics: The philosophy and the practice, Phys. Rev. Phys. Educ. Res. **13**, 020130 (2017).
[35] G. Planinšič, Project laboratory for first-year students, Eur. J. Phys. **28**, S71 (2007).
[36] European Physical Society, A Euroepan specifications for physics bachelor studies (2009), acquired May, 21, 2016 https://cdn.ymaws.com/www.eps.org/resource/resmgr/policy/eps_specification_bphys.pdf.
[37] NGSS Lead States, Next generation science standards: For states, by states (2013) acquired October, 18, 2020 https://www.nextgenscience.org/standards/standards.
[38] OECD, The future of education and skills, education 2030 (2018) acquired October, 18, 2020 http://www.oecd.org/education/2030-project/contact/.
[39] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, Am. J. Phys. **74**, 979 (2006).
[40] A. Redfors and J. Ryder, University physics students' use of models in explanations of phenomena involving interaction between metals and radiation, Int. J. Sci. Educ. **23**, 1283 (2001).
[41] Examples of web reports in Slovenian language can be found at the following address: http://projlab.fmf.uni-lj.si/arhiv/arhiv.html.
[42] V. Munier, H. Merle, and D. Brehelin, Teaching scientific measurement and uncertainty in elementary school, Int. J. Sci. Educ. **35**, 2752 (2013).
[43] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First year physics students' perceptions of the quality of experimental measurements, Int. J. Sci. Educ. **20**, 447 (1998).
[44] S. Pillay, A. Buffler, F. Lubben, and S. Allie, Effectiveness of a GUM-compliant course for teaching measurement in the introductory physics laboratory, Eur. J. Phys. **29**, 647 (2008).
[45] J. C. Arnold, K. Kremer, and J. Mayer, J. c. and, Understanding students' experiments—what kind of support do they need in inquiry tasks?, Int. J. Sci. Educ. **36**, 2719 (2014).
[46] G. Tasquier, O. Levrini, and J. Dillon, Exploring students' epistemological knowledge of models and modelling in science: results from a teaching/learning experience on climate change, Int. J. Sci. Educ. **38**, 539 (2016).
[47] J. Slisko and A. Corona Cruz, Helping students to recognize and evaluate an assumption in quantitative reasoning: A basic critical-thinking activity with marbles and electronic balance, Eur. J. Phys. Educ. **4**, 39 (2013), http://eu-journal.org/index.php/EJPE/article/view/99.

[48] I. Girault, C. d'Ham, M. Ney, E. Sanchez, and C. Wajeman, Characterizing the experimental procedure in science laboratories: A preliminary step towards students experimental design, Int. J. Sci. Educ. **34**, 825 (2012).

[49] R. Zulkarnaen, Why is mathematical modeling so difficult for students?, AIP Conf. Proc. **2021**, 060026 (2018).

[50] G. T. Prins, A. M. W. Bulte, and A. Pilot, Evaluation of a design principle for fostering students' epistemological views on models and modelling using authentic practices as contexts for learning in chemistry education, Int. J. Sci. Educ. **33**, 1539 (2011).

[51] M. Blomhøj and T. H. Kjeldsen, Teaching mathematical modelling through project work, Zentralblatt für Didaktik der Mathematik **38**, 163 (2006).

[52] T. Fuhrmann, B. Schneider, and P. Blikstein, Should students design or interact with models? using the bifocal modelling framework to investigate model construction in high school science, Int. J. Sci. Educ. **40**, 867 (2018).

[53] K. Maaß, What are modelling competencies?, Zentralblatt für Didaktik der Mathematik **38**, 113 (2006).

[54] T. Nadji and M. Lach, Assessment strategies for laboratory reports, Phys. Teach. **41**, 56 (2003).