Investigating Self-Regulated Learning as a Framework for Research-Based Assessment Feedback

Parker E. Poulos and James T. Laverty
Department of Physics, Kansas State University, Manhattan, Kansas, 66506

Bethany R. Wilcox
Department of Physics, University of Colorado Boulder, Boulder, Colorado, 80309

Assessment is an essential component to improve teaching and learning, but faculty find it difficult to improve their courses using assessment scores alone. We want to improve research-based assessment (RBA) feedback to better support faculty using RBAs to improve their courses. We take widely-recognized principles from self-regulated learning (SRL) theory that have been applied to student feedback in K-12 education, and apply them to design feedback for instructors. We conducted and analyzed interviews with faculty to determine how productive using this new RBA feedback is for instructors, and how we can further improve it. We identified five categories of change that faculty identified they would be interested in. We present these categories and discuss the feasibility of including them in future versions of our feedback.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Poulos

I. INTRODUCTION

Research-based assessments (RBAs) have a long history in physics education research, but modern RBAs are being used for wider purposes than simply as research tools to test student understanding [1]. Today, RBAs are used by researchers, instructors, and even departments to measure both student learning and instructor efficacy [2]. Developers should capitalize on this by building assessments to address the needs of instructors. To do this, developers should ensure that instructors are able to improve on their instruction using the results of RBAs, which is not always the case [2].

Feedback to an instructor from an RBA is usually given in the form of pre/post scores potentially accompanied by a statistical measure of how well students performed on the assessment [3][4]. Some RBAs allow instructors to compare their students' performance to similar classes that have taken the assessment [5]. However, these score comparisons can be difficult for instructors to make use of, especially if they don't know how to interpret their class results in the first place. These gaps in RBA feedback can leave instructors without a concrete idea of how to improve their courses based on the assessment feedback alone [2].

One way to approach filling these gaps in RBA feedback is to use design elements from self-regulated learning (SRL) theory, which will be discussed in more detail in Section II. Nicol and Macfarlane-Dick [6] identified seven principles of good feedback practice from SRL. These principles have been widely used to design feedback for students. However, the use of these principles has not yet been widely adopted in the feedback instructors receive from RBAs [2].

The Thermal and Statistical Physics Assessment (TaSPA) is an RBA designed for upper-level undergraduate thermal and statistical physics. The TaSPA allows instructors to choose from a list of 16 learning goals to assess. We also aim to provide instructors with written, actionable feedback based on their class' responses on the assessment. While this assessment provides the context in which this study is situated, the results have implications for all RBAs.

While feedback designed using SRL guidelines has historically been designed to help students regulate their learning of new material, by extension it could also help instructors to improve their courses with assessment feedback designed using these same guidelines. We have previously designed feedback for our assessment using these principles [7]. To further our understanding of the effectiveness of this method, this study seeks to answer the research questions:

1. How do instructors perceive feedback created using SRL?
2. How can we improve feedback designed with SRL to better meet the needs of instructors?

II. SELF-REGULATED LEARNING AND FEEDBACK

SRL is a process through which learners create, achieve, and evaluate their own goals for learning [8]. The SRL process begins when a learner sets a goal for themselves. For a student, this may be setting the goal of improving their scientific writing skills. Then, the learner goes through the learning process, taking in information and synthesizing it. This could look like a student attending lectures and taking notes, then revising those notes after class is over. Once the learner has done this, they then seek external feedback on their performance. This could look like a student showing a writing tutor their term paper to receive feedback. The learner then synthesizes this feedback and develops a plan to improve their performance. For a student, this may look like coming up with a plan for editing and submitting their term paper. Then, the learner implements these changes and sets new goals to begin the cycle anew [8].

In the context of the instructor as the learner, the SRL process is the same, though the execution may look different. First, an instructor will set learning goals for their course. These goals are what the instructor wants the students to learn throughout the course. This is best exemplified by learning goals receiving the preface "By the end of the course, students should be able to...." Next, the instructor teaches the course with these learning goals in mind. Finally, the instructor seeks external feedback through a research-based assessment. Based on the results of this RBA, the instructor can plan how to improve their course so that students can better meet these learning goals. To facilitate this process, feedback from RBAs should have structural features that encourage self-reflection by learners [8].

Feedback designed with SRL in mind has seven key features, as outlined by Nicol and Macfarlane-Dick [6]. The features of feedback designed with SRL that we investigated in this study are:

1. Clarify expected performance
2. Communicate the current performance
3. Provide opportunities to close the gap between expected and current performances

These three features were chosen because they are the most relevant to the structure of the feedback. For example, one of the cut features is "facilitates the development of self-assessment (reflection) in learning" [6]. This feature was cut because it is not included as a part of the structure of the feedback, but rather is included throughout. These features have also been modified slightly to clarify the centering of the instructor as the learner in this study.

Clarifying the expected performance is essentially a reminder to the learner of their initial goal. Communicating the current performance serves to show the learner *where* they can improve in order to meet their goal. Finally, providing opportunities to close the gap between expected and current performances shows learners *how* they can improve their performance in the future.

III. METHODS

The TaSPA provides the context for this study. This assessment was designed to be administered in upper-level undergraduate thermal and statistical physics courses at the end of the course (i.e., there is no pre-test). The TaSPA feedback is designed to facilitate instructor reflection and improvement of the course the instructor teaches. The feedback reports generated by the assessment were designed using the three SRL features mentioned above [7].

We interviewed faculty on the feedback developed by the TaSPA team. These interviews were an hour-long, semi-structured, and focused on instructor reactions to the feedback reports shown. The three rounds of interviews were conducted under slightly different interview protocols and versions of the presented feedback. Changes to the interview protocol were made between rounds of interviews to focus investigations on SRL components of the feedback. Feedback report versions didn't deviate from the initially mentioned SRL components. All changes to the content of the feedback were connected to results found from previous interview iterations. The changes to interview protocol and feedback versions will be mentioned where they are relevant to the results of this study.

Instructors of thermal and statistical physics were chosen on a volunteer basis to participate in these interviews. Solicitations were sent to department chairs, who were asked to forward the solicitation to instructors teaching upper-division thermal and statistical physics. Demographic information was taken for the first ten interviews, but not the subsequent five, so demographic information will not be included in this analysis. In total, we analyzed 14 interviews, 10 from Spring 2021, 3 from Spring 2023, and 1 from Spring 2024. An additional interview was conducted in Spring 2024, but was not included due to its short length (15 minutes) and lack of relevant data within the interview contents. All discussion of participants will be done through the use of pseudonyms.

In total, 177 quotes were extracted from the interview transcripts for analysis. Quotes were extracted based on whether the participant was expressing an opinion/reaction to the feedback presented, and many quotes included participants' "think-aloud" reasoning for their opinions. Thus, some quotes were much longer than others, but no discrimination was made between shorter quotes and longer ones in analysis.

The anonymized data was collected and put into structural code categories of "goals," "current state," and "advice," corresponding directly to the three features of SRL-designed feedback highlighted in the theory section. Quotes were sorted into these codes based on the section of feedback being discussed. For example, if a participant was giving their opinion on the restatement of goals in the feedback, then that quote would be coded into the "goals" category. Once the initial structural coding was completed, the data within these categories were coded again based on participant tone. These code categories were termed "positive," "negative," and "other." The "other" category contains quotes that were neither positive nor negative. These quotes were mainly instructors making sense of the feedback aloud.

In a second round of analysis, all the data was coded again, this time with special care to identify areas where participants wanted more than what the feedback gave them and areas where the participants may have been confused. All quotes were coded to look for instructor recommendations for improvements. We then performed a thematic analysis on the data following Braun and Clarke's approach [9]. We then looked at which of the SRL categories these "changes" quotes fell into in the previous analysis to see where the most changes were needed in the feedback.

A. Limitations

All but one of the feedback reports shown to instructors in these interviews were hypothetical, meaning that the reports were not for the instructor's own class. It is difficult to know whether instructors would have interpreted the feedback reports differently if they had been for their own courses. Thus, the results discussed in this paper cannot be extrapolated to determine how instructors use this feedback to improve their own course. We also acknowledge that the volunteer-based sampling method used to recruit instructors for these interviews may have caused selection bias of instructors who are already amenable to using RBAs and other research-based tools in their classrooms.

Using the TaSPA feedback as example feedback for instructors in these interviews means that these results can mainly be used to make claims about RBAs in upper-level undergraduate thermal and statistical physics, which is relevant to a narrow group of instructors and researchers. However, these results can also be extended to RBAs designed with the goal of facilitating instructor reflection and improvement, as SRL theory centers this goal and was an integral part of the feedback design process for the TaSPA.

Some of the themes found in the second analysis were mentioned by only a small number of participants. We kept these themes because we wanted to acknowledge the shared opinions of these participants and because we already had a small dataset of only 14 participants, so even two people sharing a similar opinion would be a significant fraction.

IV. RESULTS

The distribution of codes across both analyses can be seen in Table I. The "changes" quotes (in parentheses in Table I) were then sorted by type of change recommended. For example, if an instructor was discussing how they would like to see a comparison of their course to other similar courses, that recommendation would be sorted into the "course comparison" category. There were five major categories of changes that emerged, each of which is discussed in greater detail below.

Course Comparison: This category highlights the desire of instructors to compare their class' results to other classes across the country. This category had four quotes across two participants. For example, Bryant said,"I think if there's some comparison that can be made to students at other places, because I think that that's something that is really hard to determine as an individual instructor."

Feedback Specificity: This category was the largest with 28 quotes across 11 participants. Instructors in this category wanted more specific and actionable feedback. For example, Diaz said, "If you want to include more actionable things, then perhaps a bit more specific on what has helped students learn particular skills. Even if you don't want to implement them, here's a list of suggestions that people have found to be relevant for the practice."

Student Response Detail: This category was focused on instructors wanting to know exactly where students made mistakes on the assessment. This category contained 11 quotes across five participants. For example, John said, "It would be very valuable for me to know where the students are going wrong. So not just 'yeah they were not able to model this' but what are the wrong models that they are facing. So not just what did they fail to do, but in what creative ways did they fail because students are very clever. They're not just wrong. They have good reasons for being wrong."

Class Performance: This category contained four quotes across two participants and focused on instructor confusion or curiosity on the links between individual score distributions and total class performance. For example, Adams said, "I think the hard question is what is an appropriate distribution? Is it zero [percent of students] in the [not met] category? I mean, I'm curious what you all would think really."

Miscellaneous: This category contains suggestions from instructors that did not fit into any one of the above categories. This category contains four quotes across three participants. Two of these quotes were assessment content-specific suggestions which are not relevant to the current investigation. One was an instructor wondering aloud if it would be possible to get information from non-academic jobs on what skills they value from graduates. Finally, the last was an instructor wondering if it would be possible to aggregate scores from the assessment over multiple semesters to gain more statistically valid results to use in course modification.

TABLE I. Distribution of positive, negative, and other quotes in goals, current state, and advice categories. In parentheses next to each is the number of changes quotes that fell into each of these groupings. Total quote numbers should not include these changes quotes, as they are emergent from within the existing categories.

	Goals	Current State	Advice
Positive	20 (1)	13 (1)	14 (1)
Negative	8 (3)	17 (12)	37 (23)
Other	6 (1)	46 (5)	16 (3)
Total	34 (5)	76 (18)	67 (27)

V. DISCUSSION

Based on the initial coding scheme, we determined that the SRL components of feedback are a good start from which to build actionable feedback, but additional insight is needed to maximize the usefulness of the recommendations for instructors. With 20 positive quotes and eight negative quotes in the goals category, it was by far the most successful of the three SRL principles. We take this to mean no major changes are needed to the sections of the feedback that restate the goals. For the current state and advice categories, changes are needed to make these sections as useful as possible for instructors. These categories have more negative or other quotes than they have positive quotes. We argue that instructors having more negative things than positive things to say about these sections of the feedback indicates that these sections need to be investigated in more detail. This investigation follows in the rest of the discussion.

We address only the first three categories of changes here. "Class Performance," is not covered, as the TaSPA team has already discussed this topic in a past publication [10]. The "Miscellaneous" category doesn't contain enough data for us to discuss in this context.

A. Course Comparison

Instructors in this study want to compare their course's results to other courses. If we added this feature, we would most likely model its implementation on that of PhysPort, which has successfully implemented this feature for many of the assessments on their platform [11]. We would share score breakdowns for classes of similar students and allow instructors to compare their own scores to these.

We also understand that instructors may use comparison of their class to similar students to gauge how serious the problem in their course is. However, this could also be seen as a way to circumvent the process of self-reflection. If the learning goal is that "By the end of the course students should be able to...," then other instructors' results should carry no weight in an instructor's own self-reflection. Feedback designed with the SRL process in mind should have the primary goal of encouraging instructors to reflect and think critically about their course based on the information provided by their student responses. The purpose of our assessment feedback is to empower instructors to reflect on their own course design and teaching to improve their courses, and we believe that this comparison change would not align with our purposes.

B. Feedback Specificity

Instructors want the feedback to be more specific and provide more resources. This is possible to do in many ways. First, the feedback can be rewritten to focus on specific content areas students stumbled on, as we will discuss in the next subsection. The feedback may also include references to research around specific topics that students struggle with. These resources would ideally also provide teaching resources for these specific topics such as tutorials or demonstrations to be performed in class.

One reason we hesitate to include topic specific tutorials in our recommendations is due to how few of these have been studied for upper level thermal and statistical physics topics. If more research were available on these, we may be more inclined to direct instructors to these resources. As it stands, the task of rewriting all feedback to include tutorial resources for each topic covered by the assessment is too momentous a task for the TaSPA team to undertake at this time.

Another recommendation from instructors in this category was including more general teaching resources for instructors. One concern with implementing this is the volume of resources that could be given to instructors. We want instructors to feel that improving their course is possible, and inundating them with too many resources could be detrimental to this goal. Selecting which resources to give to instructors is not only a time- and labor-intensive task, but also requires explicit value judgements about what methods are most useful for instructors from a variety of different teaching backgrounds. We do not endeavor to make this kind of library of teaching methods, as it already exists in PhysPort [12].

Another concern is that telling instructors how to run their course could come off poorly both for instructors long-established in a particular style of teaching and for instructors who already use active learning methods but don't see major improvements in student mastery semester over semester. To make individualized feedback for both cases, developers would need not only information about what teaching strategies are used in the classroom, but also detailed information on how instructors are implementing these strategies. While the TaSPA does collect the former, we do not currently have plans to collect the latter in enough detail to provide individualized, teaching specific feedback. Further, even if faculty do report the active learning techniques they use, there is evidence that the self-reporting of this data is not always accurate [13]. Individualized feedback reports that target specific instructional style as well as student outcomes may be unattainable, but developers can address what sections of content instructors can improve their teaching in based on student outcomes and focus on developing feedback that will facilitate instructor self-reflection on how to improve their teaching.

C. Student Response Detail

Instructors want to know about the specific mistakes students make. One way to approach changing the feedback to provide this description for instructors is adding a breakdown of what mistakes students could have made in their reasoning, which we have implemented in the most recent iteration of TaSPA feedback. This breakdown focuses on the students who got the task partially correct to tell instructors how students missed part of a question and how they got part of it correct. There is not enough data yet to say whether this change has improved the feedback or not in the eyes of instructors, as only one interview was conducted after this change.

We have deliberately chosen not to give instructors access to the tasks of the assessment itself. We have written problems that rely extensively on real-world context for the assessment, and these real-world examples were not trivial to write. If students were to be coached in these scenarios, instructors may see scores that reflect students recalling an example rather than using their knowledge and skills to interpret a new scenario to them. The feedback currently informs instructors of student performance in terms of the topic or skill covered in the task, which aligns with the goals we have in designing the assessment itself.

VI. CONCLUSIONS AND FUTURE WORK

After analyzing the 14 interviews, we found that instructors have mixed reactions to the specific assessment feedback presented to them, and they want four broad categories of changes to the feedback developed by the TaSPA team: course comparison, feedback specificity, student response detail, and class performance. We have discussed here the ways in which three of these four changes could be implemented, along with the limitations of implementing them. Overall, we believe that the feedback from the assessment could be improved by implementing some of these changes in order to address instructor needs. However, some of the changes requested by instructors are not feasible or are in direct conflict with the goals of this assessment. These changes recommended by instructors should be taken into account by RBA developers when writing feedback for their assessments. While these changes may not all need to be made, an explanation to instructors about why developers make these choices may be necessary for instructors to effectively interpret the feedback from the assessment.

A future study could investigate whether implementing these changes desired by instructors will actually help them make productive changes to their courses. Another area of interest could be investigating these changes in the broader context of RBAs as a whole to determine whether these elements are useful for a broader range of instructors and assessments.

ACKNOWLEDGMENTS

We would like to acknowledge Amali Priyanka Jambuge for conducting ten out of the 14 interviews used in this study. We would also like to acknowledge Michael Freeman, Tyler Garcia, and Bill Bridges for for their feedback on drafts of this paper. This work was completed as part of NSF Grants #2013339 and #2013332. We would also like to acknowledge the Kansas State University Physics department for their financial support of this work.

[1] A. Madsen, S. B. McKagan, and E. C. Sayre, Best practices for administering concept inventories, The Physics Teacher **55**, 530 (2017).

[2] A. Madsen, S. B. McKagan, M. S. Martinuk, A. Bell, and E. C. Sayre, Research-based assessment affordances and constraints: Perceptions of physics faculty, Physical Review Physics Education Research **12**, 010115 (2016).

[3] A. Madsen, E. C. Sayre, and S. B. McKagan, Effect size: What is it and when and how should I use it? (2016).

[4] S. B. McKagan, E. C. Sayre, and A. Madsen, Normalized gain: What is it and when and how should I use it? (2022).

[5] B. R. Wilcox and H. Lewandowski, Students' views about the nature of experimental physics, Physical Review Physics Education Research **13**, 020110 (2017), publisher: American Physical Society.

[6] D. Nicol and D. Macfarlane-Dick, Formative assessment and self-regulated learning: a model and seven principles of good feedback practice, Studies in Higher Education **31**, 199 (2006).

[7] A. P. Jambuge, *Research-Based Assessment Design in Physics: Including Scientific Practices and Feedback for Physics Faculty*, Ph.D., Kansas State University, United States – Kansas (2021), iSBN: 9798790634109.

[8] D. L. Butler and P. H. Winne, Feedback and self-regulated learning: A theoretical synthesis, Review of Educational Research **65**, 245 (1995).

[9] V. Braun and V. Clarke, *Thematic Analysis: A Practical Guide*, 1st ed. (SAGE Publications Ltd, London ; Thousand Oaks, California, 2021).

[10] M. T. Freeman, A. Sirnoorkar, J. T. Laverty, and B. R. Wilcox, Applying Voting Theory to Mastery Grading; A Study of Faculty Interpretation of Course-Level Categorical-Score Distributions (2023) pp. 101–107, iSSN: 2377-2379.

[11] PhysPort Data Explorer.

[12] S. B. McKagan, L. E. Strubbe, L. J. Barbato, A. M. Madsen, E. C. Sayre, and B. A. Mason, PhysPort use and growth: Supporting physics teaching with research-based resources since 2011, The Physics Teacher **58**, 465 (2020), arXiv:1905.03745 [physics].

[13] D. Ebert-May, T. L. Derting, J. Hodder, J. L. Momsen, T. M. Long, and S. E. Jardeleza, What we say is not what we do: Effective evaluation of faculty professional development programs, BioScience **61**, 550 (2011), publisher: American Institute of Biological Sciences.