Applying Voting Theory to Mastery Grading; A Study of Faculty Interpretation of Course-Level Categorical-Score Distributions

M. T. Freeman,[1] Amogh Sirnoorkar,[2] James T. Laverty,[2] and Bethany R. Wilcox[1]

[1]*Department of Physics, University of Colorado, 390 UCB, Boulder, CO 80309*
[2]*Department of Physics, Kansas State University, Manhattan, Kansas 66506*

The usefulness of a research-based assessment to an instructor can vary widely depending on how student performance on the assessment is presented. Currently, the Thermal and Statistical Physics Assessment (TaSPA) is being developed with a novel reporting method to offer targeted course-improvement strategies based on student performance rather than numerical student scores. This novel reporting method, however, brings with it unique challenges with respect to characterizing course-level performance. To address these challenges, we explore voting theory as a framework to assist us in understanding the implicit value judgements in how we decide on the feedback we generate for instructors. We have also surveyed faculty perception of course-level categorical performance distributions to learn about trends and areas of consensus in how faculty interpret performance distributions, which will inform what feedback TaSPA gives instructors based on their course performance.

2023 PERC Proceedings edited by Jones, Ryan, and Pawl; Peer-reviewed, doi.org/10.1119/perc.2023.pr.Freeman

I. INTRODUCTION

To improve students' learning we must first be able to measure it. Within Physics Education Research (PER) this is primarily accomplished by research-based diagnostic tests. Considerable work is done to ensure these research-based assessments (RBAs) are reliable and valid with the goal that they can be used to improve courses by providing data on student learning directly to instructors. In practice, the onus of interpreting RBA data and deciding what to do with it is on instructors. However, making sense of the results of RBAs to inform concrete changes to classroom instruction can be challenging, even for the most practiced instructors [1]. This presents a barrier to the widespread use of these valuable tools as faculty have limited time and resources.

This work is situated within the development of a new RBA: the Thermal and Statistical Physics Assessment (TaSPA). TaSPA is currently being developed with specific attention to addressing the previously mentioned barrier by reporting results in a way that provides actionable feedback and includes suggestions for improving a faculty member's course. Fundamentally, this feedback is based on a summary of the course-level performance (i.e. a single course performance metric). However, in order to be actionable, TaSPA does not just report a course mean and standard deviation, which can be difficult to interpret. Instead, TaSPA reports a course-level, categorical distribution of how many students achieved, partially achieved, or did not achieve the learning goal (see Sec. II). Ideally, this allows faculty to more clearly interpret their students' performance [2]. This solution, while potentially allowing us to report results of RBAs in a more useful way, has its own unique challenges; when rating individual students into categorical bins (instead of a continuous score category), how do we generate feedback that meaningfully represents the course as a whole based on a distribution of categorical variables (i.e., how many students achieved, partially achieved, or did not achieve the learning goal)?

When interpreting these categorical distributions at a course level, a decision must be made about how to aggregate students' individual performance categories into course-level metrics that describe the class performance as a whole. This decision significantly influences the degree to which these metrics are interpretable for instructors. Assessment research within PER has historically avoided this concern by reporting only students "scores" (i.e., a single continuous number meant to represent their performance on the assessment as a whole). For near-continuously distributed scores with a sufficiently large sample size, we can often describe the aggregate of these individual scores by a Gaussian distribution with a mean and standard deviation, then use that model to extrapolate characteristics of the full population of interest (not all of which is in our data set). However, in practice, single-course data sets are often not large enough to justify these assumptions, and, even when they are, these aggregate metrics can be hard to interpret (e.g., what do I change when my students got a 62% on average and what does that score mean?). Additionally, this approach assumes that score is a continuous variable. In the case of TaSPA the "score" a student receives is categorical; individual students' "scores" are no longer on near-continuous scales.

This issue is not unique to TaSPA; it also arises for other fundamentally categorical schemes like Likert-style assessments as well as mastery grading [3, 4] more broadly (which also bases performance on a categorical scale of mastery or not). A problematic implication of this is that the aggregation of individual scores no longer has this clear parallel to Gaussian distributions in how to aggregate and interpret individual scores. This work seeks to address this challenge by drawing on another area – voting, which also deals with the challenge of taking individual categorical "scores" or votes and determining an appropriate overall choice meant to represent the will of the population (or, in our case, the performance of the course as a whole). In the following sections, we will lay out the specifics of TaSPA's rating system, the background and specifics of voting theory, and operationalize voting theory for TaSPA with discussion of the specific challenges and questions that must be addressed in that process.

II. BACKGROUND & MOTIVATION

TaSPA was developed with specific attention to how faculty use assessment results to inform changes to their instruction based on theories of self-regulated learning and evidence-centered design (ECD) [2, 5–7]. TaSPA evaluates students with respect to their achievement of a learning goal categorically as: "Met"(M), "Partially Met"(P), or "Not Met"(N); referred to together as MPN categories. These categories are informed not only by ECD [8] but also by criteria developed for the Next Generation Science Standards [9], i.e. 3-dimensional learning [10]. To ensure this new format is useful to faculty, we previously solicited faculty feedback via interviews. This work suggested that faculty appreciate the format in which they can see the percentage of students in each performance category and the section of suggested course changes based on the overall course performance [11].

To implement this novel feedback system for TaSPA, we must decide how to turn individual categorical performance measures into a single course-level evaluation. To do so, it is worth explicitly stating our goals as they will steer us to some solutions over others (see Sec. III). TaSPA's primary goal is to inform faculty and enable them to improve their teaching; thus, we posit that it is better to err on providing more feedback than less in cases where there is ambiguity. However, we also acknowledge that RBAs are sometimes used in an evaluative capacity (e.g., for tenure and promotion), and, thus, do not want to downplay strong performance.

As discussed in the previous section, the problem seen in interpreting TaSPA's performance categories is also seen in Likert scales that are commonly used in research-based assessment instruments [12]. Historically, these instruments dealt with this challenge by aggregating individual scores in one of three ways: averaging responses on a linear scale (e.x.

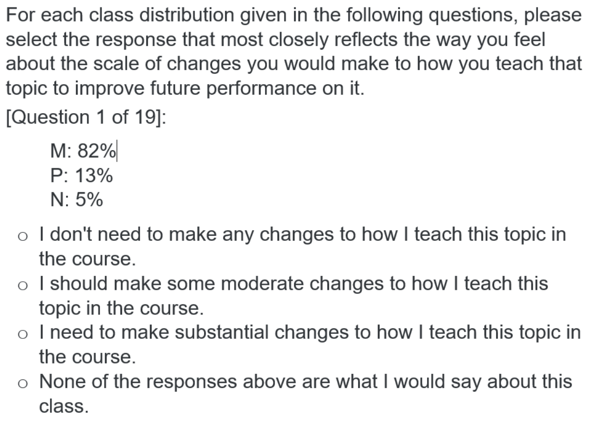
1 to 5) [13–24]; binning responses into a binary categorization (e.x. agree or disagree) [25–34]; or binning into a simple point scale and then averaging (e.x. binned point values of -1,0,1 and then averaging) [35–38]. Some notable exceptions have used more complex scoring schemes which do not fit into one of the above common categories [39–42]. The methods above, however, treat fundamentally categorical data as if it were interval when the difference between these categories is not inherently uniform between individuals or even for a single person's response [43, 44].

The sheer number of unique methods to aggregate Likert data is also an indicator that there are value judgements being made in the choice of aggregation method. An example of such a judgement for a Likert scale is that a linear scale scoring scheme values responses near the ends of the scale as discernible and uniformly removed from responses nearer to the center of the scale. This scheme places value on the extremity of a response whereas a scheme that bins the Likert scale in a binary fashion disregards the extremity and focuses only on the direction relative to the center of the scale.

III. VOTING THEORY

In many cases the value judgements being made, and how they will affect aggregation and interpretation, are implicit, which is problematic when the effects of the judgements are not consistent over the aggregation population. To make these judgements more explicit, and to provide a set of analysis tools for these decisions, we can use a special case of Social Choice Theory (a set of frameworks for the aggregation of individual metrics into collective metrics) [45] as applied to voting. This reframes our problem from one about a classroom of students to one of a population of voters. In this reframing, students become voters, and the metrics describing a class's performance become the outcomes of this vote. This is a particularly useful perspective to take because a voting system can be used to produce class performance metrics that, by construction, embody the properties (the value judgements and effects of them) inherent to that system.

Much of voting theory is beyond the scope of this paper; however, the important aspect of it for this work specifically is the way that it deals with voting systems. In voting theory, a voting system (also called an "electoral system") is a set of rules that describes how to take individual preferences and produce a group preference. These individual preferences may be a single choice or a ranked list. The group preference may be a single preference or multiple. This framework offers a lot of flexibility in how the preferences are aggregated and, as such, voting theory has some generalizable and powerful tools used for the analysis of these systems. One tool is the fairness criteria (an example of one is: if a preference has more than 50% support it should be selected as the group preference), which cannot all be satisfied for every possible vote by any electoral system (Arrow's Impossibility Theorem) [46]. This allows us to draw a parallel to implicit value judgements in how we aggregate individual student scores. By casting our aggregation methods in the framework of voting theory these fairness criteria become the value judgments that allow us to weigh the possible violations of behaviours we might desire from our aggregation method.



For each class distribution given in the following questions, please select the response that most closely reflects the way you feel about the scale of changes you would make to how you teach that topic to improve future performance on it.

[Question 1 of 19]:

M: 82%
P: 13%
N: 5%

○ I don't need to make any changes to how I teach this topic in the course.

○ I should make some moderate changes to how I teach this topic in the course.

○ I need to make substantial changes to how I teach this topic in the course.

○ None of the responses above are what I would say about this class.

FIG. 1. The first question on the survey. Each question is on its own page with a short reminder of the instructions and has a different category distribution seen in the percentages below the question's label. The first three response options will be referred to as "No Change", "Moderate Change", and "Significant Change".

IV. METHODS

This work's research questions are: how do faculty interpret TaSPA's category distributions, and what value judgments do faculty make while doing so. To answer these questions, we created and administered a survey to faculty that had taught or were teaching undergraduate courses, to analyze their responses with voting theory. The survey was designed to take 5 to 10 minutes to complete to minimize survey fatigue. The survey consisted of a preface, 19 short questions, and a demographics section. Each of the 19 questions ask the respondent to select how significantly they might change how they teach a topic in a hypothetical course based on a provided course-level category distribution. An example of this format can be seen in Fig. 1. The first 4 of the 19 questions were given to all respondents as controls, while the remaining 15 were pulled randomly from a larger set of 105 (see Sec. IV A). Faculty were asked to choose from options spanning 'making no changes' to 'making substantial changes' to their instruction, and a fourth option if they felt that their reaction to a distribution would not fit into the existing options.

The preface to the survey had the goal of putting the respondent into the mindset of someone who has just finished teaching a class where they administered a research-based diagnostic test that reported class performance broken down by the course topics. In this hypothetical situation, they have just received the results (in the format and language that TaSPA uses) for their class and must decide how significantly they might change how they teach these topics. The preface also explained the MPN categorization and explicitly addressed concerns or issues faculty might have with the survey (identified in a pilot administration of the survey). To do this we: explicitly stated the goal of the survey as “to investigate faculty interpretation of our novel reporting format”; acknowledged that teaching is complex, but noted we are asking them to imagine the distributions are the only information available to base their changes on; and acknowledged that some distributions would seem similar but were designed to probe for edge cases and nuances in interpretation.

A. Question Design

The first 4 questions of the total 19 given were the same for all respondents. This had two main purposes: to see how respondents argeed or disagreed on the same distributions, and to see how respondents interpreted different extreme distributions (e.g., distributions that are close to being entirely in one of the MPN categories). Three of these control questions (C-N, C-P, C-M in Fig. 2) were randomly generated to be in the extreme cases of high M or P or N. The last control (C-Cent.) was picked to be roughly evenly distributed in the categories as we were interested in how faculty would interpret this particular point in the distribution space.

The remaining 15 questions were pulled from 15 groups of 7 questions. Each of the 7 points (questions) in each group was centered on a randomly generated MPN distribution with a spread of 6 more points around it (see Fig. 2). This was done by taking the MPN percentages and pushing the distribution “towards” (adding 4% to one category and subtracting 2 from the others) or “away” (subtracting 4 from one category and adding 2 to the others respectively) from the extremes. The size of the spread around the central distribution was chosen to keep the percentage difference low but still meaningful. This clustering around a central distribution allowed us to get an idea of how respondents would have responded differently to small local changes in our choice of distributions as well as a sense of agreement in a region around a central distribution.

From each of the 15 groups of 7, each respondent saw one randomly selected distribution such that the 7 were equally given to respondents. The 15 central distributions were chosen as a compromise between wanting more coverage over possible distributions and survey length. These competing desires led us to 15 randomly generated center distributions by starting with approximately 20 and then removing questions that were in the most densely populated regions of the distribution space. This left us with 15 random questions spread over the possible distributions with 6 additional questions around each. These 105 total distributions were the final ones we settled on in addition to the 4 controls (see Fig. 2).

The demographics questions at the end of the survey asked for: the type of institution respondents belonged to, how many years they had been teaching at the undergraduate level, how many years since they last taught at the undergraduate level, what their typical class size was, and how often they used RBAs. These questions were chosen to provide additional insight into the background of individual respondents.

B. Survey Context

The survey was administered in the Spring of 2023 and gathered 33 responses, 9 of which were incomplete (the faculty member did not answer all questions), and 1 that was completed but all 19 questions had the 4th response option selected (“None of the responses above are what I would say about this class”). These 9 responses are not included in analysis. Of the remaining 23 faculty, only 3 ever selected the 4th response option. The maximum number of times an individual respondent selected the 4th option was 6 times, and on no question did these 3 faculty unanimously select the 4th option. Thus, the 4th option is excluded from our analysis as it did not occur often enough to support clear trends. All faculty self-identified as belonging to a “4-year Research Focused” institution and had a minimum of 6 years teaching experience. Of all faculty, 87% reported that they were currently teaching undergraduate courses, all others had taught an undergraduate course at most 2 years prior. Class size was highly variable, with 60% of respondents teaching classes of 10 to 110 students and the next largest population being faculty teaching classes typically larger than 200 students. In response to being asked to select how often they have used RBAs, faculty had the option to select “Always” (N=8), “Sometimes” (N=6), “I’ve tried them but no longer use them” (N=6), “Never” (N=1), and “I was not aware of such tests prior to this survey” (N=1). One respondent selected “Other” writing “Always if available and can be scored meaningfully.”

V. RESULTS & DISCUSSION

Responses to the survey are plotted in Fig. 2 (color mixing described in the caption). “No Change” was selected most for C-M (74%); Q18 had the next largest percentage of "No Change" (30%). Without additional questions in the region between C-M and Q18, we cannot infer whether the change in this percentage drops off suddenly or smoothly, but it is notable that even a question with a 80% Met did not get unanimous consensus from the respondents that no changes to instruction would be necessary. To identify consensus (or not), we took every group of seven points and collapsed the faculty responses for each group into a single “point” by identifying the option selected by the largest number of respondents for each distribution in that group. The collapsed “points” were then evaluated based on how many faculty agreed with this collapsed evaluation. This “agreement” is equivalent to how much consensus there is between faculty on their interpretation of distributions within each grouping. All groupings except 6 achieved an agreement of 67% (i.e., a supermajority) or more indicating general consensus amongst our respondents.

Part of the goal of this work is to inform the development of feedback for TaSPA to ensure it is consistent with faculty interpretations of our score distributions. To establish a qualitative model of faculty consensus, we have grouped the responses into 5 regions denoted by the gray dashed and solid lines seen on Fig. 2. These lines were drawn qualitatively

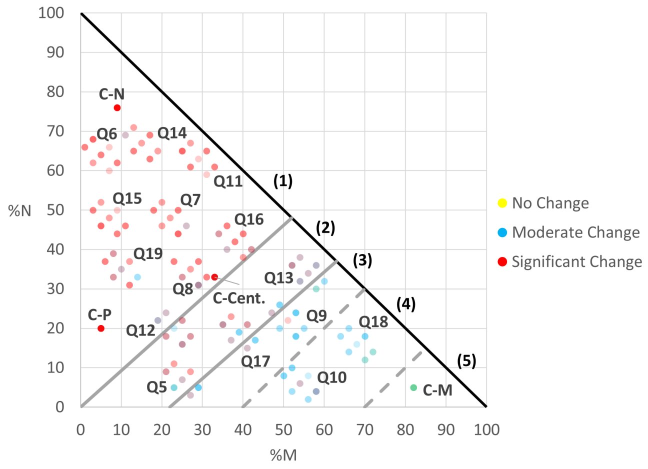


FIG. 2. Scatter plot of all distributions (plotted on %M(et) and %N(ot Met), with %P=%100−%M−%N, of the distribution the faculty responded to). Intensity of a point's color is proportional to the number of responses while hue indicates the mixing of the response categories (note, C-M is green instead of yellow due to this color mixing). The 4 control points are labeled "C-X" where "X" represents the majority category (C-Cent. is the center control). All other groups of distributions are labeled by their question number. Note that "Q8" only has 6 of the 7 points plotted due to a typographical error that resulted in no respondents seeing one of the possible distributions. Plotted in solid lines are the boundaries of the "PN" ambiguity region (see Sec. V). Plotted in dashed lines are the boundaries of the "MP" ambiguity region (see Sec. V). These 2 regions split the entire graph into 5 regions labeled "(1)" through "(5)."

so that they did not go through any points and separated the low-agreement groups (the 6 mentioned previously) from the high-agreement groups. Bounded by solid lines is the region that should contain the crossover between Moderate and Significant Change, while the dashed lines should contain the crossover between Moderate and No Change. This splits space into 5 regions termed 1 through 5 in Fig. 2; in these regions we have: agreement on Significant Change; ambiguity between Moderate and Significant Change; agreement on Moderate Change; ambiguity between No Change and Moderate Change; and agreement on No Change, respectively. Note that the large gap in points between regions 4 and 5 make the positioning of the dashed line dividing them highly uncertain, though erring on the side of more changes (or giving more feedback, see Sec. II) means putting the boundary closer to C-M. This point is discussed further in Sec. VI.

To evaluate the degree to which this simple region model fits the data, each faculty response was compared to the model and the model was scored based on number of points: correctly predicted (i.e. Significant, Moderate, and No Change in regions 1, 3, and 5 receptively); incorrectly predicted (i.e. No Change in regions 1, 2, or 3; Moderate Change in 1 or 5; or Significant Change in 3, 4, or 5); and unpredicted (i.e. any point in regions 2 or 4 that is not incorrect). The percent of points in each of these categories was calculated per respondent and then averaged. The average correct rate was 59%, the average incorrect was 16%, and the average unpredicted was 25%. The correct and incorrect percentages indicate how often the model was "right"; however, the interpretation of the unpredicted category is nuanced. It is tempting to assume we want this number to go to 0%, but that would necessarily increase the incorrect category due to the inherent variation in faculty responses. Thus, the goal of this category is to account for the variance in perception between faculty of these distributions in our model such that we can decrease the number of incorrect responses without over-fitting our data.

VI. CONCLUSIONS & FUTURE WORK

Here, we explored faculty interpretations of a novel approach to reporting student performance on research-based assessments. Based on their responses, we created a simple region model to convert categorical distributions of student performance to course-scale feedback that can help faculty decide whether to make changes to their instruction. As a preliminary study with small N, this simple region model provides a useful tool for understanding faculty interpretations to inform development of useful feedback. In particular, this work will guide us in developing the algorithm that generates feedback reports for TaSPA based on class MPN distribution, ensuring feedback aligns with faculty interpretation. This work also allows for investigation of how faculty might interpret these categorical distributions and allows us, as assessment developers, to gain a better understanding of our audience and where our priorities and values might meaningfully differ. For example, the general trends in our data support the claim made earlier that, in general, faculty will prefer to make changes over not making changes, even for classes with performance levels an assessment developer might consider quite high (e.g., >70%). This supports the goal that TaSPA sets of erring on the side of providing more feedback to instructors.

While this work was motivated as part of the broader TaSPA project, because this survey was not specific to TaSPA, this procedure could be used to investigate aggregation of other categorical scales, such as mastery grading and Likert-style. Future work for this project could include developing a modified survey with more questions centered around the regions of ambiguity identified in this round and targeting a larger faculty population. This increased statistical power would allow us to identify existing voting systems that predict faculty responses. Identifying an appropriate voting theory would provide a way to look for inherent value judgments faculty, as a collective, are making when interpreting categorical scoring such as mastery grading or Likert-style questions.

ACKNOWLEDGMENTS

Thank you to Sara Moore, who introduced me to voting theory, and the TaSPA team who provided feedback on this project. This work was supported by the National Science Foundation under Grant No.s 2013332 & 2013339.

[1] Adrian Madsen, Sarah B. McKagan, Mathew Sandy Martinuk, Alexander Bell, and Eleanor C. Sayre, "Research-based assessment affordances and constraints: Perceptions of physics faculty," Phys. Rev. Phys. Educ. Res. **12**, 010115 (2016).

[2] Amali Priyanka Jambuge, Katherine Rainey, Bethany Wilcox, and James Laverty, "Assessment feedback: A tool to promote scientific practices in upper-division," in *Physics Education Research Conference 2020*, PER Conference (Virtual Conference, 2020) pp. 234–239.

[3] Silvia Heubach and Sharona Krinsky, "Implementing mastery-based grading at scale in introductory statistics," PRIMUS **30**, 1054–1070 (2020), https://doi.org/10.1080/10511970.2019.1700576.

[4] Andrew A. Cooper, "Techniques grading: Mastery grading for proofs courses," PRIMUS **30**, 1071–1086 (2020), https://doi.org/10.1080/10511970.2020.1733151.

[5] Katherine Rainey, Amali Priyanka Jambuge, James Laverty, and Bethany Wilcox, "Developing coupled, multiple-response assessment items addressing scientific practices," in *Physics Education Research Conference 2020*, PER Conference (Virtual Conference, 2020) pp. 418–423.

[6] James Laverty, Bethany Wilcox, Amali Jambuge, Katherine Rainey, and Amogh Sirnoorkar, "Development of a Thermal and Statistical Physics Assessment," in *APS April Meeting Abstracts*, APS Meeting Abstracts, Vol. 2021 (2021) p. Y15.004.

[7] Katherine D. Rainey, Michael Vignal, and Bethany R. Wilcox, "Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage," Phys. Rev. Phys. Educ. Res. **16**, 020113 (2020).

[8] Christopher J. Harris, Joseph S. Krajcik, James W. Pellegrino, and Angela Haydel DeBarger, "Designing knowledge-in-use assessments to promote deeper learning," Educational Measurement: Issues and Practice **38**, 53–67 (2019), https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12253.

[9] "Next generation science standards," [Accessed on 27 April 2023].

[10] James T Laverty, Sonia M Underwood, Rebecca L Matz, Lynmarie A Posey, Justin H Carmel, Marcos D Caballero, Cori L Fata-Hartley, Diane Ebert-May, Sarah E Jardeleza, and Melanie M Cooper, "Characterizing college science assessments: The three-dimensional learning assessment protocol," PloS one **11**, e0162333 (2016).

[11] Amali P. Jambuge, *Research-Based Assessment Design in Physics: Including Scientific Practices and Feedback for Physics Faculty*, Ph.D. thesis (2021).

[12] "Physport," [Accessed on 08 April 2023].

[13] Tim Weston, Sandra Laursen, Anne-Barrie Hunter, and Heather Thiry, "Undergraduate research student self-assessment," (2015).

[14] Anne-Barrie Hunter, Timothy Weston, Sandra Laursen, and Heather Thiry, "Urssa: Evaluating student gains from undergraduate research in the sciences," CUR Quarterly **29**, 15–19 (2009).

[15] Christine Lindstrøm and Manjula D. Sharma, "Physics goal orientation survey," (2010).

[16] Christine LindstrÃžm and Manjula Sharma, "Development of a physics goal orientation survey," Int. J. Innov. Sci. Math. Educ. **18**, 10–20 (2010).

[17] Elif Ince, Hilal Çağap, and Yasemin Deneri, "Motivation scale towards physics learning," (2020).

[18] Elif Ince, Hilal Çağap, and Yasemin Deneri, "Development and validation of motivation scale towards physics learning," International Journal of Physics and Chemistry Education **12**, 61–74 (2020).

[19] Heidi Fencl and Karen Scheel, "Sources of self-efficacy in science courses - physics," (2005).

[20] Heidi Fencl and Karen Scheel, "Research and teaching: Engaging students - an examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course," J. Coll. Sci. Teaching **35**, 20–24 (2005).

[21] Christine Lindstrøm and Manjula Sharma, "Physics self-efficacy questionnaire," (2011).

[22] Christine Lindstrøm and Manjula Sharma, "Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter?" Int. J. Innov. Sci. Math. Educ. **19**, 1–19 (2011).

[23] Kimberly A. Shaw, "Self-efficacy in physics," (2003).

[24] Kimberly Shaw, "The development of a physics self-efficacy instrument for use in the introductory classroom," in *Physics Education Research Conference 2003*, PER Conference, Vol. 720 (Madison, WI, 2003) pp. 137–140.

[25] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C.E. Wieman, "Colorado learning attitudes about science survey," (2006).

[26] Wendy Adams, Katherine Perkins, Noah S. Podolefsky, Michael Dubson, Noah Finkelstein, and Carl Wieman, "New instrument for measuring student beliefs about physics and learning physics: The colorado learning attitudes about science survey," Phys. Rev. ST Phys. Educ. Res. **2** (2006).

[27] E. F. Redish, J. M. Saul, and R. N. Steinberg, "Maryland physics expectation survey," (1998).

[28] Edward F. Redish, Jeffrey Saul, and Richard Steinberg, "Student expectations in introductory physics," Am. J. Phys. **66**, 212–224 (1998).

[29] Wendy K. Adams, Richard L. III Pearson, and Savannah L. Logan, "Perceptions of teaching as a profession for higher education survey," (2020).

[30] Richard L. Pearson III, Savannah L. Logan, and Wendy Adams, "Faculty perception insights obtained from faculty interviews during the development of the perceptions of teaching as a profession in higher education (ptap.he) instrument," in *Physics Education Research Conference 2020*, PER Conference (Virtual Conference, 2020) pp. 394–399.

[31] Wendy K. Adams, Taylor Plantt, Heather Taffe, and Monica Plisch, "Perceptions of teaching as a profession," (2020).

[32] Savannah L. Logan, Jared B. Breakall, Richard L. Pearson III, and Wendy Adams, "College faculty support for grade 7-12 teaching careers: survey results and comparisons to student perceptions," in *Physics Education Research Conference 2020*, PER Conference (Virtual Conference, 2020) pp. 291–296.

[33] Karen Cummings, Stephanie Lockwood, and Jeff Marx, "Attitudes toward problem solving survey," (2003).

[34] Karen Cummings, Stephanie Lockwood, and Jeffrey Marx, "Attitudes toward problem solving as predictors of student success," in *Physics Education Research Conference 2003*, PER Conference, Vol. 720 (Madison, WI, 2003) pp. 133–136.

[35] Ben Zwickl and Heather Lewandowski, "Colorado learning attitudes about science survey for experimental physics," (2012).

[36] Benjamin Zwickl, Noah Finkelstein, and Heather J. Lewandowski, "Development and validation of the colorado learning attitudes about science survey for experimental physics," in *Physics Education Research Conference 2012*, PER Conference, Vol. 1513 (Philadelphia, PA, 2012) pp. 442–445.

[37] Andrew Mason and Chandralekha Singh, "Attitudes and approaches to problem solving suvey," (2010).

[38] Andrew Mason and Chandralekha Singh, "Surveying college introductory physics students' attitudes and approaches to problem solving," Eur. J. Phys. **37**, 055704 (2016).

[39] Andrew Elby, John Frederiksen, Christina Schwarz, and Barbra White, "Epistemological beliefs assessment for physical sciences," (2001).

[40] Andrew Elby, "Helping physics students learn how to learn," Am. J. Phys. **69**, S54–S64 (2001).

[41] Sufen Chen, "Views on science and education questionnaire," (2006).

[42] Sufen Chen, "Development of an instrument to assess views on nature of science and attitudes toward teaching science," Sci. Educ. **90**, 803–819 (2006).

[43] Matthew Lovelace and Peggy Brickman, "Best practices for measuring students' attitudes toward learning science," CBE-Life Sciences Education **12**, 606–617 (2013), pMID: 24297288, https://doi.org/10.1187/cbe.12-11-0197.

[44] Ibrahim A Halloun, "Student views about science," Lebanon: Educational Research Centre, Lebanese University (2001).

[45] Amartya Sen, "Social choice," in *The New Palgrave Dictionary of Economics* (Palgrave Macmillan UK, London, 2017) pp. 1–20.

[46] Kenneth J. Arrow, "A difficulty in the concept of social welfare," Journal of Political Economy **58**, 328–346 (1950), https://doi.org/10.1086/256963.