Criteria for collapsing rating scale responses: A case study of the CLASS

Ben Van Dusen
Department of Science Education, California State University Chico, Chico, CA, 95929, USA

Jayson Nissen
J. M. Nissen Consulting, Corvallis, OR, 97333, USA

Assessments of students' attitudes and beliefs often rely on questions with rating scales that ask students the extent to which they agree or disagree with a statement. Unlike traditional physics problems with a single correct answer, rating scale questions often have a spectrum of 5 or more responses, none of which are correct. Researchers have found that responses on rating scale items can generally be treated as continuous and that unless there is good evidence to do otherwise, response categories should not be collapsed [1–3]. We discuss two potential reasons for collapsing response categories (lack of use and redundancy) and how to empirically test for them. To illustrate these methods, we use them on the Colorado Learning Attitudes about Science Survey. We found that students used all the response categories on the CLASS but that three of them were potentially redundant. This led us to conclude that the CLASS should be scored on a 5-point or 3-point scale, rather than the 2-point scale recommended by the instrument developers [4]. More broadly, we recommend the judicious use of data manipulations when scoring assessments and retaining all response categories unless there is a strong rational for collapsing them.

I. INTRODUCTION

Research-based assessments have played a pivotal role driving physics education research (PER) and physics course transformations [5]. These assessments have been developed to measure different areas of interest in learning physics, including content knowledge, scientific reasoning, and attitudes/beliefs about physics [6]. While the strength of the validation argument for each instrument varies, a defining feature of research-based assessments is the research that has gone into assessing the reliability and validity of them. As part of the ongoing validation process, it is common for researchers to investigation the reliability of using the instrument across contexts, student populations, and time. These investigations have typically focused on student interviews [7], question wording [8], factor analyses, and measures of reliability [9, 10] but have not as commonly examined question scoring. Unlike assessments that measure content knowledge and offer responses that are correct or incorrect, assessments that measure attitudes or beliefs typically use rating scales that offer 5 responses, ranging from *Strongly Disagree* (*SD*) to *Strongly Agree* (*SA*). The lack of a single correct answer has led to a variety of scoring styles being adopted. For example, the Sources of Self-Efficacy in Science Courses - Physics (SOSESC-P) uses a 5-point scoring scale [11], the Colorado Learning Attitudes about Science Survey for Experimental Physics uses a 3-point scoring scale (E-CLASS) [12], and the Colorado Learning Attitudes about Science Survey (CLASS) uses a pair of 2-point scoring scales [4]. We examined two empirical criteria for deciding to collapse rating scale response categories prior to scoring, lack of use and redundancy. To demonstrate the methods for testing the criteria, we apply them to the CLASS and examine the impact of using a 5-point versus 2-point scoring scheme. We conclude with general recommendations for scoring of rating scale assessments.

II. LITERATURE REVIEW

A. Scoring rating scale items

Research-based assessments typically use two different types of items: multiple-choice items with single correct answers and rating scale items with no correct answer. While PER researchers have investigated learning progressions through incorrect responses on multiple-choice items [13–15], it is generally agreed upon that multiple-choice items should be scored dichotomously as either correct or incorrect. Rating scale items offer more possible scoring methods. For example, rating scale items with 5 possible responses (e.g., *Strongly Disagree* (*SD*), *Disagree* (*D*), *Nuetral* (*N*), *Agree* (*A*), and *Strongly Agree* (*SA*)) are commonly scored in different ways [1, 4, 12, 16], such as using a 5-point scale, a 3-point scale, or a 2-point scale. Fig. 1 shows some common ways that categories on a rating scale with 5 responses are collapsed to create either a 3-point scale or a 2-point scale.

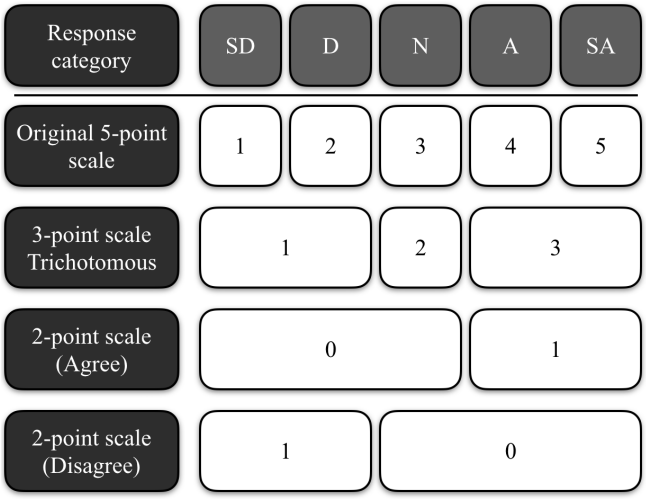

FIG. 1. Common response collapsing schemes to transform a 5-point scale to 3-point or 2-point scales. The 2-point scales shown are recommended by the developers of the CLASS [4].

A common critique of how rating scale items are scored is that the data is ordinal, not interval, and the distance between two ordered responses (e.g., *SD* and *D*) may not be equal to the distance between two other ordered responses (e.g., *N* and *A*) [17, 18]. Treating rating scale data as ordinal data would require nonparametric analyses [19]. Researchers have found, however, that when examining multiple rating scale items in aggregate the data can be treated as continuous without introducing bias [1–3]. Further, responses to individual items with at least 5 response options can generally be treated as continuous values [20, 21]. While parametric tests are often appropriate for analyzing rating scale data, it is still important to ensure that the data meets the test assumptions (e.g., independence of observations, homogeneity of variance, and a normal distribution) [22].

As each response category provides information, it is generally recommended that response categories be retained unless there is good reason for them to be collapsed [16, 23]. There are situations, however, that warrant the collapsing of response categories. One justification for collapsing a response category is that it is not being selected and is therefore not providing any information. A second justification for collapsing a response category is that it is being interpreted the same as another response category and is therefore providing redundant information. Both justifications can be empirically examined by calculating the selection rates for each response category and each response category's correlation with the overall score (i.e., point-biserial correlations). Response categories that provide additional information and should not be collapsed are those that are selected regularly and have point-biserial correlations that are ordered lowest to highest across the least to most expert-like responses. In this paper, we ex-

amine these two criteria.

B. CLASS

The CLASS was developed by Adams *et al.* [4] to assess student beliefs about physics and learning physics. Researchers and instructors have largely adopted the 2-point scoring method recommended by Adams *et al.* [4]. While Adams *et al.* [4] often used 2-point scoring methods to analyze CLASS responses, they also used the traditional 5-point scoring method in their development process. For example, they used a 5-point scale in their exploratory factor analysis. Adams *et al.* [4] recommended using a pair of 2-point scales (as illustrated in Fig. 1) when scoring the CLASS. The first 2-point scale they recommend using examines the fraction of students that agree with expert views by collapsing *SD*, *D*, and *N* then *A* and *SA*. The second 2-point scale they recommend using examines the fraction of students that disagree with expert views (or agree with novice views) by collapsing *SD* and *D* then *N*, *A*, and *SA*. Adams *et al.* [4] justified the collapsing of the *N* category because their interviews showed that students used the *N* category inconsistently. They state that the *N* category data should be removed and therefore all other response categories should be treated as being ordinal. They justified the collapsing of *SD* with *D* and *A* with *SA* because there was not complete agreement in student interpretation of the categories (i.e., two students with the same level of agreement might select different responses). It is difficult to assess the strength of the evidence to support their decision to collapse response categories because neither the data from the interviews nor any quantitative evidence was included in the publication to support the decision.

Since its publication, researchers have reexamined the validity argument of the CLASS. Sawtelle *et al.* [7] found that students at Florida International University (a Hispanic-Serving Institution) interpreted the questions as they were designed, which expanded the validation argument to include more diverse student populations. Douglas *et al.* [10], however, examined the psychometric properties of the CLASS and came to different conclusions than Adams *et al.* [4] on how to interpret CLASS data. Douglas *et al.* [10] critiqued the novel methods for factor analysis that Adams *et al.* [4] used for violating assumptions of reliability and validity. In their reanalysis of the CLASS using both exploratory and confirmatory factor analysis, Douglas *et al.* [10] conclude that the data only support the scoring of 15 of the 41 items on the instrument and that they measure 3, rather than 8, constructs. Heredia and Lewis [9] came to similar conclusions in their analysis of the Chemistry version of the CLASS [24]. We tested the full and abbreviated versions of the CLASS and found similar results for both. In this publication, we followed Douglas *et al.* [10]'s recommendation and focused on the 15 item version.

TABLE I. Data breakdown before after filtering

	Pre-filter	Post-filter
Institutions	27	26
Courses	187	185
Students	10466	8851

III. EXAMPLE - CLASS

A. Methods

Our study analyzed student CLASS data from the Learning About STEM Student Outcomes (LASSO) platform [25]. The LASSO platform is hosted on the LA Alliance website [26] and is a free resource for instructors that collects large-scale, multi-institution data by hosting, administering, scoring, and analyzing research-based assessments online. LASSO-using instructors teach in diverse institutional settings and tend to use collaborative-learning activities in their courses [27].

To clean the data, we removed assessment scores from students who answered the CLASS's filter question incorrectly or took less than 3 minutes to complete the assessment. This led to the removal of 1615 students from the dataset who did not have either a pretest or posttest score. The final dataset included data from 8851 students in 185 courses at 26 institutions (as shown in Table I).

In analyzing the student responses, we inverted the answers for any question in which SA aligned with novice views. We only used the 15 items Douglas *et al.* [10] recommended analyzing in their critique of the CLASS. For the sake of thoroughness, we repeated our analysis using the 36 items recommended by Adams *et al.* [4]. Our findings were similar using either version of the assessment, so for brevity's sake we will only discuss the 15-item version in the findings. We calculated the proportion of students selecting each response and point-biserial correlations between each response on an item with the overall score on the assessment using the *itemanalysis* package [28] in R. To calculate the overall score for this analysis; we used the same 5-category scoring method as Adams *et al.* [4] in their psychometric evaluation of the CLASS in which the range of answers from *SD* to *SA* were assigned a value of 1-5. In our comparison of the impact of different scoring methods, we score each assessment using both the 5-category method and the 2-category method recommended by Adams *et al.* [4] in which answers of *SD* to *N* are scored as a 0 and *A* and *SA* are scored as a 1.

B. Findings

Our analysis of response frequencies found that all available responses were used by students on the CLASS. Table II shows that students were more likely to agree (*A* or *SA*)

TABLE II. Response category selection proportions.

	SD (%)	D (%)	N (%)	A (%)	SA (%)
Item 1	10.7	18.4	24.1	28.3	18.5
Item 2	11.3	21.6	28.5	28.6	10.0
Item 3	7.9	15.0	26.5	31.5	19.1
Item 4	10.9	21.3	23.4	31.4	13.0
Item 5	8.0	32.7	32.7	21.5	5.1
Item 6	1.9	7.9	16.9	43.4	30.0
Item 7	1.2	4.5	14.1	41.8	38.3
Item 8	14.4	14.1	30.9	26.5	14.1
Item 9	2.6	6.9	22.8	40.9	26.8
Item 10	1.6	5.3	13.0	33.2	46.9
Item 11	1.5	4.4	18.2	42.1	33.8
Item 12	2.4	6.8	19.3	41.4	30.1
Item 13	4.9	14.1	35.3	36.2	9.4
Item 14	6.2	14.7	22.0	36.9	20.2
Item 15	5.2	11.3	22.7	38.9	22.0
Mean	6.1	13.3	23.4	34.8	22.5

TABLE III. Point-biserial correlations.

	SD	D	N	A	SA
Item 1	-0.361	-0.211	-0.065	0.173	0.369
Item 2	-0.221	-0.169	-0.043	0.203	0.225
Item 3	-0.334	-0.237	-0.118	0.179	0.365
Item 4	-0.228	-0.199	-0.110	0.196	0.321
Item 5	-0.129	-0.161	-0.044	0.200	0.222
Item 6	-0.106	-0.178	-0.166	0.006	0.265
Item 7	-0.122	-0.125	-0.188	-0.037	0.252
Item 8	-0.435	-0.200	-0.039	0.250	0.374
Item 9	-0.252	-0.250	-0.233	0.054	0.395
Item 10	-0.121	-0.154	-0.198	-0.082	0.310
Item 11	-0.232	-0.244	-0.304	-0.049	0.463
Item 12	-0.176	-0.192	-0.250	-0.004	0.383
Item 13	-0.267	-0.255	-0.110	0.245	0.279
Item 14	-0.243	-0.179	-0.124	0.080	0.337
Item 15	-0.245	-0.259	-0.207	0.117	0.400
Mean	-0.231	-0.201	-0.147	0.102	0.331

than disagree (*D* or *SD*) with an item. The most commonly selected response was *A* (34.8%) and the least commonly selected response was *SD* (6.1%). While the proportion of the time that *SD* was selected was small for some items (e.g., Item 7 - 1.2%) it was more substantial for other items (e.g., Item 8 - 14.4%).

Our analysis of correlations between responses and overall scores found inconsistent shifts in the point-biserial correlations across response categories (Table III). Some items have regular and meaningful shifts in point-biserial values. For example, on item 8, as student responses move from *SD* to *SA* their correlations with overall expert-like views of physics get meaningfully more positive. Other items, however, have small or inverted shifts in in point-biserial values. For example, on item 12, as student responses move from *SD* to *N* their correlations with overall expert-like views of physics get slightly more negative before getting more positive when moving from *N* to *SA*. Figure 2 shows the point-biserial values for each response on each item. On average, the shifts in point-biserial values from *SD* to *D* and *N* are small (0.030 and 0.054) while the shifts to *A* and *SA* are larger (0.249 and 0.228). The only items with point-biserial values that are similar (difference < 0.030) for *A* and *SA* are items 2, 5, and 13. The rest of the items have point-biserial values for *A* and *SA* that differ by 0.124 to 0.512.

Our comparison of overall scores using the 5-category versus 2-category methods and of scoring responses found that scoring student responses using 5 categories led to slightly higher pretest and posttest means than scoring with 2 categories (Table IV). The gains from pretest to posttest were similarly small and negative using either scoring style. Scoring with 5 categories, however produced smaller standard deviations and subsequently a larger effect size than scoring with 2 categories.

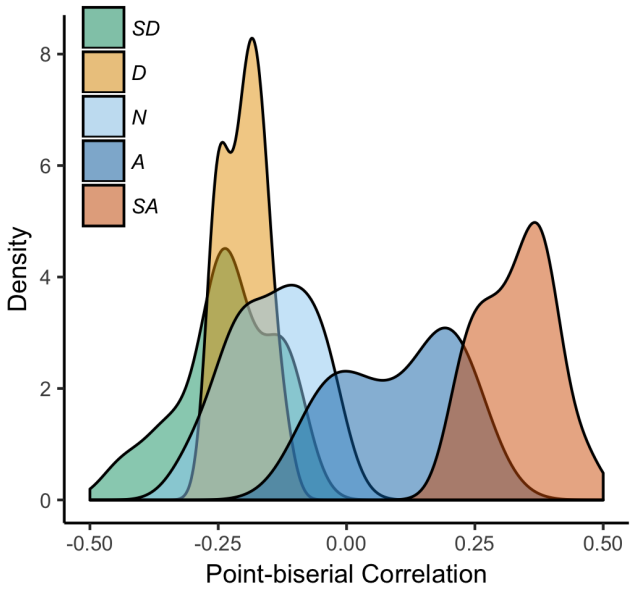

FIG. 2. Density plot of point-biserial correlations.

C. Interpretation

Examining whether our findings support collapsing categories on the CLASS, we considered two potential justifications: (1) lack of response selection and (2) similarity in correlations for response selections with overall scores (i.e., point-biserial values). The low proportion of students selecting *SD* ($< 2\%$) for some items on the CLASS (e.g., Item 6, 7, 10, and 11) make it a potential candidate for collapsing it with *D*. However, since *SD* was selected more frequently ($> 10\%$) on other items (e.g., 1, 2, 4, and 8) its frequency of selection does not justify collapsing with *D*.

TABLE IV. Student outcomes by scoring style.

Scoring style	Pretest (%)		Posttest (%)		Gain (%)	
	Mean	S.D.	Mean	S.D.	Mean	*d*
5-point	64.3	11.0	62.7	12.1	-1.6	-0.139
2-point	58.8	17.9	57.1	19.2	-1.7	-0.087

The mean point-biserial values for *SD*, *D*, and *N* only have small differences and indicate that they may be candidates for collapsing. The mean point-biserial values for *A* and *SA*, however, are meaningfully different from each other and indicate that the two responses are distinct from each other and should not be collapsed. These findings do not support using either 2-point scoring methods recommended by Adams *et al.* [4]. Our findings support analyzing CLASS data using either a 5-point scoring system or a 3-point scoring system that collapses the *SD*, *D*, and *N* categories and leaves *A* and *SA* as their own distinct categories.

The absolute shifts from pretest to posttest on the CLASS have historically been small [29] and the use of a 5-point versus 2-point scoring system does not appear to change this. The interpretation of the gains, however, could be impacted by the use of a 5-point scoring system. The cohen's *d* effect size was more than 50% larger when calculated using 5-point versus 2-point scoring. The increase in the effect size is driven by a decrease in the standard deviation when using the 5-point scoring. We caution drawing any conclusions that are too large about these differences, however, as the effect sizes are small for both the 5-point (-0.139) and the 2-point scoring (-0.087).

IV. DISCUSSION

Research-based assessments provide instructors and researchers simple methods for measuring changes in students' performances in a range of areas, including content knowledge and attitudes. To support valid and reliable claims using data from these assessments, it is important that the instruments be developed, implemented, and analyzed using high-quality methods. In the analysis of data, it is often useful to manipulate data (e.g., removing spurious data). However, as manipulations have the potential to introduce bias into data, they should be limited to only those that have strong evidence to support their use. As such, when examining rating scale data, response categories should not be collapsed without compelling evidence. If each response category is being used and is not being used redundantly, then collapsing response categories removes information and may bias results.

In our examination of the CLASS we found that while all response categories were used, the point-biserial correlation values for *SD*, *D*, and *N* were roughly equivalent, indicating that students were using all three response categories to mean similar things. This led us to recommend that the CLASS be scored using either a non-manipulated 5-point scale or collapse *SD*, *D*, and *N* categories to make a 3-point scale. The developers of the CLASS [4] recommend collapsing the responses to a pair of 2-point scales based on their interviews not identifying consistent differences in students use of SD and D or A and SA. As none of their interview data is included in their publications, it is difficult to judge the strength of their conclusion. Our findings indicate that students are using A and SA to mean different things which does not support the use of a 2-point scale. More broadly, we hope that instrument developers and users will be thoughtful in how they score assessments and consider following the methods we demonstrated in our analysis of the CLASS.

The instrument evaluation methods discussed in this paper are only small subset of the instrument evaluation methods developed by researchers in other disciplines (e.g., data science and psychometrics). We strongly recommend that instrument developers explore the space of classical test theory and item response theory to help improve assessment reliability and validity [30–32]. These tools can also test for potential biases in assessments across student demographics and courses contexts [33, 34].

V. ACKNOWLEDGEMENTS

This work is funded in part by NSF-IUSE Grant No. DUE-1525338 and is Contribution No. LAA-064 of the Learning Assistant Alliance. We are grateful to the Learning Assistant Program at the University of Colorado Boulder for establishing the foundation for LASSO and LASSO studies. We acknowledge and are mindful that the paper was primarily written at CSU Chico which stands on lands that were originally occupied by the first people of this area, and we recognize the Mechoopda and their distinctive spiritual relationship with this land and the waters that run through campus. We are humbled that CSU Chico resides upon sacred lands that once sustained the Mechoopda people for centuries.

[1] S. E. Harpe, How to analyze likert and other rating scale data, Currents in Pharmacy Teaching and Learning **7**, 836 (2015).
[2] A. R. Baggaleyand A. L. Hull, The effect of nonlinear transformations on a likert scale, Evaluation & the health professions **6**, 483 (1983).
[3] S. Labovitz, The assignment of numbers to rank order categories, American sociological review , 515 (1970).
[4] W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Physical Re-

view Special Topics - Physics Education Research **2**, 1 (2006).
[5] N. R. Council *et al.*, *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering* (National Academies Press, 2012).
[6] A. A. of Physics Teachers, *Physport* (2019), https://www.physport.org/assessments/ [Accessed: 04/30/2019].
[7] V. Sawtelle, E. Brewe, and L. Kramer, Validation study of the colorado learning attitudes about science survey at a hispanic-serving institution, Physical Review Special Topics-Physics Education Research **5**, 023101 (2009).
[8] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Physical Review Special Topics-Physics Education Research **3**, 010102 (2007).
[9] K. Herediaand J. E. Lewis, A psychometric evaluation of the colorado learning attitudes about science survey for use in chemistry, Journal of Chemical Education **89**, 436 (2012).
[10] K. Douglas, M. Yale, D. Bennett, M. Haugan, and L. Bryan, Evaluation of colorado learning attitudes about science survey, Physical Review Special Topics-Physics Education Research **10**, 020128 (2014).
[11] H. Fencland K. 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, Journal of College Science Teaching **35**, 20 (2005).
[12] B. R. Wilcoxand H. Lewandowski, Studentsâ epistemologies about experimental physics: Validating the colorado learning attitudes about science survey for experimental physics, Physical Review Physics Education Research **12**, 010123 (2016).
[13] T. Smith, K. Gray, K. Louis, B. Ricci, and N. Wright, Showing the dynamics of student thinking as measured by the fmce, in *Proceedings, Physics Education Research Conference 2017* (2018) p. 380.
[14] K. Louis, B. Ricci, and T. Smith, Determining hierarchies of correctness through student transitions on the fmce, in *Proceedings, Physics Education Research Conference 2018* (2019).
[15] Á. Pérez-Lemonche, B. C. Drury, and D. Pritchard, Mining student misconceptions from pre-and post-test data., International Educational Data Mining Society (2018).
[16] H. Jeongand W. Lee, The level of collapse we are allowed: Comparison of different response scales in safety attitudes questionnaire, Biometrics & Biostatistics International Journal **4**, 100 (2016).
[17] S. Jamieson *et al.*, Likert scales: how to (ab) use them, Medical education **38**, 1217 (2004).
[18] G. D. Armstrong, Parametric statistics and ordinal data: A pervasive misconception, Nursing Research **30**, 60 (1981).
[19] L. Cohen, L. Manion, and K. Morrison, *Research methods in education* (routledge, 2002).
[20] T.-C. Hsuand L. S. Feldt, The effect of limitations on the number of criterion score values on the significance level of the f-test, American Educational Research Journal **6**, 515 (1969).
[21] B. M. Byrne, *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (Routledge, 2016).
[22] B. Van Dusenand J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, Physical Review Physics Education Research **15**, 020108 (2019).
[23] P. Grimbeek, F. Bryer, W. Beamish, M. D'Netto, *et al.*, Use of data collapsing strategies to identify latent variables in chp questionnaire data, Stimulating the'Action'as Participants in Participatory Research: Volume 2 , 125 (2005).
[24] W. K. Adams, C. E. Wieman, K. K. Perkins, and J. Barbera, Modifying and validating the colorado learning attitudes about science survey for use in chemistry, Journal of Chemical Education **85**, 1435 (2008).
[25] B. Van Dusen, Lasso: A new tool to support instructors and researchers, American Physics Society: Forum on Education **Fall**, 12 (2018).
[26] L. A. Alliance, *Learning About STEM Student Outcomes (LASSO) Platform* (2019), https://learningassistantalliance.org/ [Accessed: 04/30/2019].
[27] X. Herrera, J. Nissen, and B. Van Dusen, Student outcomes across collaborative-learning environments, in *Proceedings, Physics Education Research Conference 2018* (2019).
[28] C. Zopluogu, Package 'itemanalysis', cran.r-project.org (2018).
[29] A. Madsen, S. B. McKagan, and E. C. Sayre, How physics instruction impacts studentsâ beliefs about learning physics: A meta-analysis of 24 studies, Physical Review Special Topics-Physics Education Research **11**, 010115 (2015).
[30] C. Zabriskieand J. Stewart, Multidimensional item response theory and the conceptual survey of electricity and magnetism, Physical Review Physics Education Research **15**, 020107 (2019).
[31] L. Ding, Theoretical perspectives of quantitative physics education research, Physical Review Physics Education Research **15**, 020101 (2019).
[32] M. Planinic, W. J. Boone, A. Susac, and L. Ivanjek, Rasch analysis in physics education research: Why measurement matters, Physical Review Physics Education Research **15**, 020111 (2019).
[33] R. Henderson, J. Stewart, and A. Traxler, Partitioning the gender gap in physics conceptual inventories: Force concept inventory, force and motion conceptual evaluation, and conceptual survey of electricity and magnetism, Physical Review Physics Education Research **15**, 010131 (2019).
[34] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the force concept inventory with modified module analysis, arXiv preprint arXiv:1905.06176 (2019).