Evaluation of high school student responses to the Colorado Learning Attitudes about Science Survey

Julian S. Martins and William E. Lindsay

University of Colorado, School of Education, 2445 Kittredge Loop Dr, Boulder, Colorado 80305, USA

(Received 26 October 2020; revised 8 December 2021; accepted 7 February 2022; published 25 April 2022)

This study describes a psychometric evaluation of high school student responses to the Colorado Learning Attitudes about Science Survey (CLASS), and the subsequent development of a more parsimonious attitudes and beliefs survey structure for potential use with K-12 students in physics courses aligned with the Next Generation Science Standards. Pre- and postsurvey response data were obtained from high school students in the 2017–2018 and 2018–2019 school years, whose instructors were partnered with the *Physics through Evidence: Empowerment through Reasoning* project. Exploratory factor analysis methods were used to propose a physics attitudes and beliefs survey with a more parsimonious factor structure, and confirmatory factor analysis methods provide support for the survey's structural stability. These preliminary results suggest that the CLASS is fertile ground for the development of shorter attitudes and beliefs surveys that may be more easily implemented and interpreted by instructors in K-12 contexts. Replication analyses and potential uses of the more parsimonious survey structure are also discussed.

DOI: 10.1103/PhysRevPhysEducRes.18.010132

I. INTRODUCTION

For more than a century, efforts seeking to reform physics education in the United States have attempted to shift general instruction from solely didactic methods toward approaches that seek to actively engage students in inducing physics principles from evidence [1]. The results of these reform efforts have been mixed, highlighting the historical difficulty within the United States to implement and sustain novel and research-based educational practices on a national scale [2]. One common theme of these reform movements is a concern for fostering positive attitudes and beliefs of students toward the practice of science, as well as toward the relationship between the scientific enterprise and their personal lives. Since the late 1800's, reform educators have called for fostering a "spirit of science" or a "spirit of inquiry" in students, and similar language has been used throughout the decades to describe how appreciating the beauty and internal coherence of the scientific process should be one of the main foci of science education [1]. More recently, studies have connected student attitudes and beliefs, as well as perceptions of self-efficacy regarding physics, to both a lack of diversity within the field and retention rates in the university context [3,4].

Many survey instruments with the purpose of attempting to measure student attitudes and beliefs toward science have been developed to help address these concerns. Two examples of survey instruments in the context of physics are the Maryland Physics Expectations Survey [5] and the Colorado Learning Attitudes about Science Survey (CLASS) [6]. The CLASS in particular is classified as a highly reliable instrument [7], and it has been adapted for use in other subject-matter disciplines such as biology [8], chemistry [9], and laboratory-focused physics courses [10]. Within the U.S. university context, many studies have been published that report CLASS results in a wide variety of university-level introductory physics courses, such as calculus-based physics courses for science majors [11], inquiry-based conceptual courses directed at elementary education majors [12], and modeling-instruction courses [13], among others [14]. The CLASS has also been translated into multiple languages and applied in non-U.S. contexts [15,16].

However, few studies exist that involve applications of the CLASS with high school students, and those that do have occurred mostly outside of the U.S. [17,18]. Considering that many students' first contact with physics occurs in K-12, the relatively small number of studies involving the application of attitudinal surveys with students in this context offers an opportunity to enrich the academic debate surrounding students' attitudes and beliefs toward science. Attending to the development of positive

TABLE I. Examples of CLASS statements by category.

Agreement indicates expertlike beliefs	Statements without consistent expert opinions, not scored	Statements where agreement indicates novicelike beliefs
2. When I am solving a physics problem, I try to decide what would be a reasonable value for the answer.	4. It is useful for me to do lots and lots of problems when learning physics.	6. Knowledge in physics consists of many disconnected topics.
28. Learning physics changes my ideas about how the world works.	9. I find that reading the text in detail is a good way for me to learn physics.	18. There could be two different correct values to a physics problem if I use two different approaches.
38. It is possible to explain physics ideas without mathematical formulas.	41. It is possible for physicists to carefully perform the same experiment and get two very different results that are both correct.	35. The subject of physics has little relation to what I experience in the real world.

student attitudes and beliefs toward physics prior to the college experience becomes more significant when one considers studies showing how CLASS results with undergraduate students tend to remain static, or even decrease, throughout their time in college [19]. While some teaching methods and curricular approaches have been reported to positively impact undergraduate student attitudes and beliefs as measured through the CLASS [13,20], courses with these characteristics tend to be offered to nonscience majors, and the traditionally oriented courses offered to science majors are commonly shown to have a negative attitudinal impact, or no impact at all [14]. Thus, we argue that fostering positive student attitudes and beliefs toward physics should be a goal of physics instruction prior to the college experience, and within this objective, the CLASS provides a means of investigating the potential attitudinal impacts of different pedagogies and curricular approaches.

This study is concerned with investigating the psychometric properties of high school student response data to the physics CLASS, collected by instructors partnered with the *Physics through Evidence, Empowerment through Reasoning* project. Through descriptive statistics, exploratory factor analysis, and confirmatory factor analysis, responses were analyzed to inform the extraction of a more parsimonious survey instrument inspired by the full CLASS. A shorter survey instrument with replicable structure and a greater ease of statistical interpretation has the potential to facilitate comparison between results of future studies investigating high school student attitudes and beliefs toward physics. Thus, the research questions guiding this study are

1. *What are the psychometric properties of CLASS results with a high school student sample population?*
2. *Do CLASS responses obtained from a high school student sample population inform the development of a shorter physics attitudinal survey instrument?*
3. *Does a smaller factor structure obtained from students' beginning-of-year CLASS responses provide acceptable fit indices when applied to the same students' end-of-year CLASS responses?*

II. THE CLASS

The CLASS is an instrument that seeks to "probe students' beliefs about physics" and measure the extent to which students have beliefs that align with those held by professional physicists [6] (p. 1). Influenced by prior science-related attitudinal surveys such as the Maryland Physics Expectations Survey [5] and the Epistemological Beliefs Assessment about Physical Science [21], the CLASS is designed to be applicable with a wide variety of student populations. The instrument seeks to address a broad variety of issues considered significant by educators in the process of learning physics, and it was designed to be easily administered, with item wordings that facilitate adaptation for use in other science disciplines [8,22]. Since their inception, the CLASS and its variations have been widely used in science education research [14].

The CLASS has 41 attitudinal statements about physics, to which students indicate their level of agreement on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree," as well as one "filter" question to identify students that are not reading statements prior to responding. Some item statements are worded so that agreement with them indicates "expertlike thinking" (e.g., alignment with beliefs held by expert physicists), while agreement with other statements indicate "novicelike thinking" (see Table I). CLASS results are scored under an ordinal method, which does not assume an exact mathematical relationship between each level on the scale being used [23]. In other words, it is not assumed that the difference between "agree" and strongly agree corresponds to the same difference between agree and "neutral" regarding the psychometric construct measured by the instrument. The CLASS developers defended the use of a 5-point scale during application, as cognitive interviews with students revealed a meaningful distinction between levels of

TABLE II. Categories and statement groupings in CLASS scoring.

Category or statement grouping	Category or grouping description	Statements in category
Personal interest	Do students feel a personal interest in or connection to physics	3, 11, 14, 25, 28, 30
Real world connection	Seeing the connection between physics and real life	28, 30, 35, 37
Problem solving (general)	No description	13, 15, 16, 25, 26, 34, 40, 42
Problem solving (confidence)	No description	15, 16, 34, 40
Problem solving (sophistication)	No description	5, 21, 22, 25, 34, 40
Sense making or effort	For me (the student), exerting the effort needed toward sense making is worthwhile	11, 23, 24, 32, 36, 39, 42
Conceptual understanding	Understanding that physics is coherent and is about making sense, drawing connections, and reasoning, not memorizing or making sense of math	1, 5, 6, 13, 21, 32
Applied conceptual understanding	Understanding and applying a conceptual approach and reasoning in problem solving, not memorizing or following problem solving recipes	1, 5, 6, 8, 21, 22, 40
Statements not categorized, but with "consistent expert perspective" (included in "overall" scoring)	No description	2, 10, 12, 17, 18, 19, 20, 27, 29, 38
Learning style questions (not a validated category in terms of expertlike responses)	What students believe to be useful for learning	4, 9, 12, 16, 19, 33
Statements not scored	Described as "on slate for revision"	7, 41

agreement or disagreement during statement interpretation. However, interviews also showed a lack of consistency within student interpretations of response options, in the sense that different respondents with the same conviction of belief could respond differently to the same statement [6]. Under this rationale, neutral responses are disregarded for scoring purposes, and CLASS responses are then collapsed into a 2-point scale that treats strongly agree and agree as the same answer for expertlike statements, while the opposite occurs for novicelike statements.

When scored, CLASS results for a sample population are represented in relation to various dimensions of student attitudes and beliefs toward physics. Most of the survey statements are grouped into eight categories, developed through an iterative process combining statistical data-driven approaches and predetermined categorizations proposed by physics experts and teachers. These categories are not fully independent of each other, some statements are not scored in relation to any given category, and some statements are not scored at all. Nonscored statements are kept in the survey because they provide formative assessment information that is potentially valuable for physics faculty. Therefore, in the automatic scoring spreadsheet for the CLASS [7], results for a given sample set of pre- and postassessment responses are shown in various ways (see Table II):

i. "Overall," which takes into account the 36 scored statements with "consistent expert responses."
ii. For each of the eight category groupings, which are various combinations of 26 statements.
iii. In relation to statement groupings that lack consistent expert perspectives but are still seen to provide formative assessment information of use to teachers.

III. LITERATURE ON THE CLASS AND K-12 PRACTICAL MEASURES

A. CLASS results with traditional instruction

Studies involving the CLASS instrument generally apply it in the form of a pre- and postsurvey, where students take the survey at the beginning and end of a course experience. The differences between student results in the pre- and postsurvey ("shifts") are interpreted in relation to various course characteristics and valued outcomes, such as reform structures [11], curricular or pedagogical approaches [12,13], and student conceptual learning [24]. Commonly reported results from most of these investigations show few positive shifts toward expertlike thinking among student populations. For instance, in their meta-analysis of student beliefs about learning physics, Madsen *et al.* [14] indicated that the most common effect of lecture-based instruction was an overall regression of student beliefs toward being more aligned with "novice" views of science. In connection to this finding, researchers have correlated traditional instructional practices with negative shifts in student CLASS results [25]. Research

into student perceptions of self-efficacy has proposed that traditional physics instruction may not support students in generating positive self-efficacy states [26], and that this phenomenon could be related to the lack of student diversity in a field whose population has been historically dominated by white males.

B. Differences in CLASS results across demographics

Studies have also investigated students' interpretation of CLASS statements and the patterns that may exist between student demographics and survey results. Gray *et al.* [27] found that students understand what the beliefs of expert physicists are but do not personally agree with them, suggesting that the importance of developing expertlike beliefs was not being fostered in their course experiences. CLASS scores have been shown to vary in relation to gender and race in the U.S. and abroad, with white male students often having the most expertlike scores in pre- and postsurvey applications [15,18,26,28,29]. These results lend themselves to some interesting interpretations, given that the norms of modern professional science are influenced by the norms of European academic societies whose membership has historically been composed mainly of white men. Potential explanations aside, given the calls for increased gendered and racial diversity within the physics community [30], gendered and racially biased results such as these indicate the need for increased attention toward the fostering of positive attitudes and beliefs in marginalized student groups.

C. The CLASS as a predictor for student learning and retention

Studies have revealed a potential connection between student beliefs and conceptual understanding, showing not only that the beliefs held by students prior to instruction can be a predictor for the conceptual gains that they will achieve [25], but also that students without a certain threshold of expert beliefs are unlikely to become a physics major [31]. Furthermore, longitudinal studies investigating student beliefs related to physics show that these beliefs remain essentially static throughout the entire undergraduate experience [17,31]. To summarize, the attitudes and beliefs toward physics that students in a science major hold when they leave university may well be the same—or even more novicelike—than the attitudes and beliefs they had when they entered. This leads to an uncomfortable view of current higher education in physics: undergraduate courses do not invite students into the field. Instead, they perhaps filter out students who do not *already enter* the university context with the expertlike beliefs that are required to succeed. A similar threshold idea regarding student perceptions of self-efficacy has also been proposed in other research regarding retention rates in university introductory physics courses [4].

D. Contexts for positive CLASS results

Not all studies involving applications of the CLASS in undergraduate populations report negative or nonexistent shifts in survey results. When instructors either explicitly attend to fostering expertlike beliefs about physics with their students or implement reform structures, positive CLASS shifts are reported in lecture-based courses with large class sizes directed at science majors [11]. Moreover, some curricular approaches and course structures consistently show positive impacts on CLASS results—all share the characteristic of engaging students in the process of constructing and defending conceptual models of physics phenomena [14]. For instance, Otero and Gray [12] reported positive CLASS shifts in *Physics and Everyday Thinking* (PET), an inquiry-based course for nonscience majors. In PET, students inductively formalize physics concepts through guided-inquiry experiments, consensus discussions, and engagement in science practices [32]. Other examples of studies reporting positive CLASS shifts include implementations of modeling instruction in undergraduate physics courses [13] and courses applying the *Physics by Inquiry* curriculum [33]. While these curricular approaches differ in the extent that they explicitly address the nature of physics learning, scientific epistemology, and the history of science, they share the characteristic of providing students the opportunity to learn physics by engaging in practices that align with those of professional scientists.

E. Psychometric properties of the CLASS

Finally, a recent line of inquiry into the CLASS focuses on analyzing the various psychometric qualities of data gathered by the CLASS and its variants. Van Dusen and Nissen [29] used original CLASS response data collected through the Learning About STEM Student Outcomes platform to discuss criteria for collapsing rating scale responses. They argued for the use of a 3-point or 5-point scoring scale for the CLASS, instead of the 2-point scale recommended by Adams *et al.* [6], stating that collapsing response categories could remove information and bias interpretations of survey results. Heredia and Lewis [22] used factor analysis methods to perform an evaluation of response data for the chemistry version of the CLASS, proposing that the survey provides fertile ground for the development of shorter instruments with reasonable psychometric properties, and further suggesting that similar analyses be performed for data obtained with the physics and biology versions of the CLASS. In line with this recommendation, Douglas *et al.* [34] used factor analysis methods to conduct an evaluation of the physics CLASS, arguing for a shorter version of the instrument with a more parsimonious structure. They proposed a 15-item version of the CLASS with three factors. This factor structure was similar to the one proposed by Heredia and Lewis [22] for the chemistry CLASS. However, the demographics of the

sample population in Douglas *et al.* [34], composed largely of white male students, was a potential cause for concern in relation to the generalizability and replicability of their proposed factor structure with other student populations. Lastly, Cahill *et al.* [35] implemented the CLASS in their multiyear evaluation of an interactive-engagement curriculum, and used factor analysis methods to develop a pair of orthogonal factors using 25 CLASS items that most closely captured aspects of students' approaches to learning and problem solving.

The developers of the original CLASS responded to Douglas *et al.* [34], critiquing its methodology and findings [36]. They argued that nuances in student beliefs were lost with a reduced factor structure, and that using a subset of the original survey would "substantially reduce its utility" (p. 10). For example, Douglas *et al.* [34] suggested collapsing the *"personal interest"* and *"real-world connection"* factors from the original CLASS into a single category named *"personal application and relation to the world"* (p. 7). Wieman and Adams [36] responded to this suggestion by stating that collapsing these factors would reduce the sensitivity of the instrument, as instructors would no longer be able to differentiate between shifts in *personal interest* which had been correlated strongly with becoming a physics major, and shifts in *real-world connection* which had a more negative shift than *personal interest* in the population originally tested by the CLASS developers [6]. Therefore, it could be the case that removing CLASS statements that do not correlate well with other statements would lead to the loss of important information about student perceptions. To justify these recommendations, the CLASS developers emphasized the purpose and utility of the CLASS as a formative assessment. While parsimony may make for a tidier analysis and more replicable results across population samples, a shorter instrument would also provide less information about student attitudes and beliefs. That said, Wieman and Adams [36] agreed that the validity of a survey instrument is not a "one-time stamp," and that evidence toward reliability and validity should be collected for different populations (p. 9).

F. Practical measures for use in K-12 contexts

By grounding their argument against condensing the CLASS in the utility of the instrument for providing nuanced formative information to teachers, the original developers assumed that teachers would have the motivation, time, and expertise to implement, score, analyze, and apply the information provided by the survey to inform instruction. Since Wieman and Adams [36] argue that the CLASS is to be used for formative purposes, it follows that the instrument is intended to be implemented and analyzed while teaching a physics course. Indeed, Wieman and Adams [36] state that the utility of the CLASS lies in "defining particular aspects to which instruction can be targeted" (p. 4). For instructors to use the CLASS in this manner, their students must take the CLASS and have their responses properly scored and analyzed, a task which may not be feasible for K-12 instructors during an academic year. Although the CLASS scoring spreadsheet automatically scores student data and generates graphical results, this output must still be interpreted by instructors and given that the CLASS has not been analyzed psychometrically with high school populations, this interpretation is not a trivial task.

We argue that it is important to acknowledge how the individual and contextual differences between university and K-12 instructors may influence instrument implementation and perceived utility. The above capacity assumptions may hold true for university instructors seeking to implement the CLASS, many of whom have received quantitative research training through their graduate studies, have less teaching responsibilities than K-12 teachers, and have access to university-based resources such as the learning about STEM student outcomes platform. This may result in more time to score and interpret the results of the CLASS, alongside having less student surveys to analyze. More support may be required for similar instrument usage to occur in the K-12 context. Studies of data use by K-12 teachers have indicated the importance of explicitly attending to the capacity of individuals and their local communities to support the productive implementation of formative instruments like the CLASS. At the individual level, teachers have varying degrees of knowledge regarding how data can be used to inform instruction, a conceptual relationship referred to as pedagogical data literacy [37]. In their article on the growing need for data-driven decision making in education, Mandinach and Gummer [37] discuss how pre- and in-service educators should be exposed to multiple experiences throughout their careers to develop data literacy. They argue that institutions of teacher preparation must take a more active role in developing educator data literacy, given that educators are increasingly expected to use diverse forms of data to inform instruction. An important component of pedagogical data literacy is making sense of the link between assessment results and their implications for specific practices [38], and research has found that effective data use is encouraged by collaboration, as distributed expertise is shared across teachers [39]. Local organizational conditions such as structured time for analysis, timely access to evidence of student learning or affective outcomes, tools and guides for collaboration, and sustained professional development, have been found to enable data use [38,40,41].

Unfortunately, most studies have pointed to a lack of data use capacity in K-12 systems at the individual [42] and organizational levels [43]. Some reasons identified for this lack of capacity include a dearth of professional development opportunities, limited courses on data usage during preservice training, the decoupling of organizational

policies and practice, and few opportunities for productive social networking around data use practices [42,43]. Scholars concerned with improving instructional practices in K-12 settings have attempted to address data use capacity issues by proposing the development and usage of formative instruments known as "practical measures" [44]. The intention of a practical measure is to provide evidence about the impact of specific educational practices that is gathered and interpreted in the context of those practices. In their paper describing the design of a system of practical measures (one of which was a short student survey), Penuel *et al.* [45] state that they should be embedded in teaching practices, predictive of valued outcomes, able to generate data that can be used to improve practice, and frequently usable by educators. The usability of practical measures is described in terms of dimensions such as learnability and efficiency, which relate to how easily the measure can be used for the first time and to how quickly the measure's tasks can be performed [45]. To ensure that practical measures are taken up by K-12 teachers with limited capacity for data use, key design features should also include relative ease of analysis, where analysis results have clear implications for instruction.

In contrast to the reasoning proposed by Wieman and Adams [36] for keeping the CLASS complex, proponents of practical measures contend that such complexity often discourages instrument usage, as teachers are overwhelmed with a deluge of information that does not offer clear guidance for shifts in practice [44,46]. In addition, practical measures are especially important for teachers with a lack of individual, network, or organizational capacity for understanding or responding to the results of complex instruments [47]. The previously reviewed research suggests that this is the existing condition for most K-12 physics teachers. Following this reality, the development and validation of a shorter, more practical, and more easily interpreted version of the CLASS may be a useful step for providing formative evidence to K-12 teachers that can be used to inform practice. This conclusion aligns with prior research into the psychometric properties of the CLASS. For instance, Heredia and Lewis [22] also argued for the increased feasibility of shorter surveys, particularly in contexts with strong time constraints where there are concerns that survey respondents may not fully complete a longer instrument.

It should be stated that this study is not intended to criticize the CLASS instrument and argue for large-scale revisions to the instrument for all purposes. Rather, the objective is to explore the psychometric properties of response data for a new sample population of interest and use this data to develop a more parsimonious and statistically interpretable set of statements, whose properties can be potentially replicated in other high school contexts. As contended above, high school physics teachers are much more likely to use research-based instruments like the CLASS to inform their instruction if the instrument is easily implemented and interpreted. Therefore, this study has an overarching goal of supporting future lines of research into the impacts of innovative curricular approaches in high school physics education, alongside offering a practical measure to K-12 teachers that requires less capacity for usage. To produce this measure, we first calculated descriptive statistics and performed exploratory factor analysis on CLASS preresponses to produce four potential factor structures. We then tested these potential factor structures and the structures proposed by Douglas *et al.* [34] and Cahill *et al.* [35] on CLASS postresponses to identify the structure that best fit our sample of data. Finally, we employed bootstrapping methods to provide further evidence about the structural validity of the final proposed instrument. These methods are described in further detail in the following sections.

IV. METHODS

A. PEER Physics

The student response data analyzed in this study were gathered by high school instructors partnered with the *Physics through Evidence, Empowerment through Reasoning* (PEER Physics) project at the University of Colorado Boulder. PEER Physics is a suite that provides teachers, schools, and districts with curricular materials and sustained professional development. These resources help promote inquiry-based physics instruction aligned with the pedagogy proposed by *A Framework for K-12 Science Education* [48] and upheld in the *Next Generation Science Standards* [49]. The PEER Physics approach is adapted from *Physics and Everyday Thinking–High School* (PET-HS) [50,51], which is in turn adapted from the PET undergraduate course for nonscience majors [32]. PEER Physics thus shares the main characteristics of PET of inviting students into the process of experimentation, building scientific models of physics phenomena, and inducing physics principles from data and the process of consensus. PEER Physics uses the methodology of placing students in small groups that routinely engage in whole-class discussions to make sense of experimental observations and conceptual reasoning, as is also the case with other similar curricula, such as *Physics by Inquiry* and *Modeling Instruction* [13,33].

High school teachers partnered with PEER Physics engage in sustained professional development with a focus on understanding and implementing NGSS ideals, through immersion activities that apply the same pedagogical approaches as PEER Physics activities themselves. As instructors gain more experience with PEER Physics, their professional development evolves to include teacher-directed inquiry into common problems of practice, as well as training to become professional development

TABLE III. Missing response information.

2017–2018						
Number of possible	Number of PRE missing	%	% of PRE missing in last 10 items	Number of POST missing	%	% of POST missing in last 10 items
7790	33	0.4	54.5	38	0.5	52.6
2018–2019						
Number of possible	Number of PRE missing	%	% of PRE missing in last 10 items	Number of POST missing	%	% of POST missing in last 10 items
12 546	52	0.4	88.5	50	0.4	86

providers for their school districts. Since the original PET-HS curriculum started being field-tested in Colorado in 2014, a body of literature has been developed that supports the impact of the curricular suite on increasing student equity [52], student conceptual understanding [53], and professional learning [54].

B. Data collection and student demographics

A total of 15 instructors participated in data collection during the 2017–2018 and 2018–2019 academic years, with four instructors participating in both years. Participating instructors teach at public and charter high schools in the Midwest and Pacific Northwest of the United States. Initial application of the CLASS instrument for predata collection occurred in September of each academic year, while postdata collection occurred in May. The survey was given in paper and pencil format. During pre- and postapplications of the CLASS instrument, participating students also completed a physics diagnostic assessment developed by the PEER Physics team which included a short survey component. Participating teachers were asked to implement both the CLASS and the PEER Physics diagnostic assessment so that the impacts of PEER Physics instruction could be investigated through multiple data sources.

The data collected for this study represents approximately 800 students attending 11 high schools in seven school districts. Many students neglected to provide names on pre- or postresponse sheets, making it difficult to provide more precise estimates regarding numbers of individual students impacted by PEER Physics participation. Unlike prior investigations involving the CLASS in undergraduate semester-course contexts, this study uses data collected over an entire academic year with younger student populations, which brings unique challenges and constraints regarding the collection and treatment of matched data. For instance, in some districts it is common for students to be shifted between instructors after an academic semester, and therefore some students wind up with missing pre- or postdata. Another factor affecting data collection was the lack of explicit incentives for students to provide full responses to the CLASS. Thus, overall response rates were low.

Given that matched response sets to both the PEER Physics diagnostic and the CLASS instrument were desired, student information was kept only if a set of conditions were simultaneously met:

i. Student had the same instructor for the entire academic year;
ii. Student provided responses to >80% of the PEER Physics diagnostic assessment in pre- and postapplication;
iii. Student provided responses to >80% of the CLASS in pre- and postapplication.

This initial filtering of data reduced the full sample set to 683 students (261 students in 2017–2018, and 422 students in 2018–2019). In a following stage of student response data selection, two additional conditions were applied:

iv. Student provided responses to >80% of PEER Physics demographic survey questions in pre- and postapplication;
v. Student correctly answered the CLASS filter question in pre- and postapplication.

After these two stages of treatment, the sample set was reduced to 497 students. Missing response data for the reduced sample set can be seen in Table III.

As the CLASS instrument has 41 attitudinal statements and the reduced dataset included 497 individuals, the maximum number of possible responses was 20 377. Incomplete responses were analyzed to investigate missingness, and the reduced sample set contained 85 missing responses in the pre-survey and 88 in the postsurvey, a missingness rate for total responses of 0.4% in both academic years. An initial analysis of the mechanism for missing responses indicated that survey fatigue within sample respondents could have affected response rates. A higher-than-expected proportion of missing responses were concentrated in the final 10 items of the CLASS survey, particularly in 2018–2019. To avoid biased results due to this nonrandom source of missingness, particularly within the final 10 items in the original CLASS, students with missing responses were also removed from the dataset. This resulted in a final sample set of 423 students. More

TABLE IV. Student sample demographics for both academic years.

2017–2018											
School information		Gender			Race						
School [Number of students] (Number of instructors)	% Free or reduced-fee lunch	% Female	% Male	% NA	% African American	% Asian	% Caucasian	% Latino/a	% Native American	% Pacific Islander	% NA
1 [34] (1)	49	94.1	2.9	2.9	2.9	5.9	35.3	55.9	0	0	0
2 [28] (1)	32.8	39.3	60.7	0	0	3.6	64.3	32.1	0	0	0
3 [31] (1)	32	22.6	77.4	0	0	3.2	77.4	16.1	3.2	0	0
4 [10] (1)	18	20	80	0	0	0	100	0	0	0	0
5 [13] (1)	45	46.2	53.8	0	0	0	15.4	76.9	0	0	7.7
6 [48] (3)	40.2	47.9	52.1	0	2.1	4.2	47.9	41.7	2.1	2.1	0
2018–2019											
School information		Gender			Race						
School [Number of students] (Number of instructors)	% Free or reduced-fee lunch	% Female	% Male	% NA	% African American	% Asian	% Caucasian	% Latino/a	% Native American	% Pacific Islander	% NA
1 [16] (1)	49	100	0	0	12.5	6.3	56.3	25	0	0	0
2 [74] (3)	40.2	60.8	39.2	0	1.4	1.4	66.2	14.9	1.4	0	14.9
3 [37] (1)	32	21.6	78.4	0	5.4	8.1	67.6	16.2	0	2.7	0
4 [39] (1)	58	51.3	48.7	0	5.1	2.6	12.8	76.9	0	2.6	0
5 [34] (1)	44.5	38.2	61.8	0	0	14.7	58.8	23.5	2.9	0	0
6 [26] (2)	88.9	57.7	38.5	3.8	3.8	15.4	15.4	65.4	0	0	0
7 [33] (1)	18	54.5	45.5	0	0	3	93.9	0	3	0	0

Overall (both academic years)										
School information	Gender			Race						
...	% Female	% Male	% NA	% African American	% Asian	% Caucasian	% Latino/a	% Native American	% Pacific Islander	% NA
...	51.5	47.8	0.7	2.4	5.2	57.4	30.3	1.2	0.7	2.8

TABLE V. Item removal iterations informed by descriptive statistics.

Item Number	Lower, raw, upper estimates of α	Average interitem correlation	CLASS items with corrected item-total correlation <0.3
36	0.78, 0.81, 0.83	0.11	1, 2, 8, 12, 18, 19, 21, 22, 27, 38
26	0.83, 0.85, 0.87	0.18	6, 10, 17
23	0.83, 0.85, 0.87	0.2	⋯

information on sample demographics can be seen in Table IV.

C. Data analysis

This study's analyses were primarily conducted using *R Studio*, except for the CFA analysis of CLASS postresponse data, which used *Mplus 7.4*. Prior to analysis, items 4, 7, 9, 33, and 41 were removed. These items do not have consistent expert responses, and are thus considered neutral items that are not used for scoring purposes in the original CLASS. Items 1, 5, 6, 8, 10, 12, 13, 17, 18, 20, 21, 22, 23, 27, 29, 32, 35, and 40 are worded so that agreement indicates alignment with novice beliefs toward physics, these items were reverse scored using the "*recoder*" R package. The "*psych*" R package was used to calculate descriptive statistics for the remaining 36 CLASS items, obtain estimates of internal consistency, perform EFA, and perform replication or bootstrapping analyses.

V. FINDINGS

A. Item descriptive statistics and internal consistency

Average scores for the remaining 36 CLASS items ranged from 1.85 to 4.12, with standard deviation values from 0.81 to 1.16. Sixteen items had a median of 4, 18 items had a median of 3, and only two items had a median of 2, indicating an overall positive skew to response distributions. Minimum and maximum values for every CLASS item were 1 and 5, respectively, indicating that every possible response option was selected by respondents. Only two items, 12 and 24, were found to have skew or kurtosis with a magnitude larger than 1, indicating that their response distributions were concentrated around certain values—of these, item 12 was also one of the two items to have a median of 2. The relatively high medians among most of the items in the dataset, in combination with standard deviations close to 1.0 for all items, suggests non-normality of the data. However, for ordinal data of this nature, an assumption of normality would be tenuous even if descriptive statistics seemed to support it.

Reliability estimates were obtained for the 36 non-neutral CLASS items. Cronbach's α is a commonly reported statistic of internal consistency for survey instruments, and it is a function of average interitem correlation and the number of items in the analyzed instrument. Cronbach's α for the entire scale was calculated as 0.81, with 95% confidence boundaries of 0.78 and 0.83. However, Cronbach's α on its own is misleading as a measure of internal reliability, as the average inter-item correlation of the 36 CLASS items in this dataset was 0.11. Given that α is a function of average interitem correlations and the number of items, this suggests that the calculated value of α in this case is skewed by the large number of items in the scale. Therefore, item-whole correlations (corrected for item overlap) were calculated for each item, and this statistic was used as a criterion to inform the removal of further items. After each iterative removal of items, Cronbach's α was recalculated for the remaining items in the scale.

In the literature, cutoffs for item-total correlations vary from 0.3 to 0.5 [55,56]. Although it is desired that the final reduced survey instrument is able to measure an overall single latent construct, the original CLASS was designed as a multifactor instrument and thus it is assumed that the reduced survey will also contain multiple correlated factors. Therefore, a lower cutoff of 0.3 for corrected item-total correlation was chosen in this case. With this condition for a first filtering process, items were iteratively removed from the dataset until all remaining items possessed corrected-item total correlations above 0.3. Table V provides more information regarding this first iterative item-removal process, including reliability estimates for all remaining items at each stage.

B. Exploratory factor analysis

Exploratory factor analysis (EFA) was chosen as the statistical tool to inform the development of a shorter physics attitudinal survey from the CLASS. EFA examines the "pairwise relationships between individual variables" in a scale and seeks to extract latent factors from them [57]. It is thus a data reduction tool which serves to aid in psychometric analyses of a scale. EFA is well suited for the overall objective of this study, which is to investigate whether high school student CLASS item response patterns can be modeled through a smaller set of latent variables. Three separate factor structures were obtained through EFA for this data, and the steps in this procedure are outlined below. Factor extraction and rotation methods are discussed first, followed by factor retention decisions, interpretation of results from different generated factor structures, and finally an evaluation of robustness.

1. Methods of factor extraction and rotation

Communalities were initially estimated through each item's squared multiple correlation [58]. In a second stage of item filtering, one item with SMC below 0.2 was removed from the dataset prior to factor extraction. Multiple methods of factor extraction exist, and in this case, principal axes factoring [57,58] was chosen as it was not tenable to hold an assumption of multivariate normality within the dataset. During extraction, factors were rotated to help interpretation. Factor loadings for items can be conceptualized as axis coordinates, where each axis represents a factor. With N extracted factors, each item possesses N factor loadings that characterize its location within an N-dimensional space. Put simply, when factors are rotated, the axes that represent the factors are shifted so that items cluster closer to the axis they load more strongly upon. Ideally, this leads to nondominant factor loadings being placed near zero, which then facilitates the interpretation of factor loading matrices.

Rotations are classified as orthogonal or oblique, depending on whether underlying factors are uncorrelated [57] or correlated [59], respectively. In this case, given that the original CLASS was designed as a multi-factor instrument, and that in the social sciences it is generally expected that factors within a single scale correlate with each other [57], an oblique rotation was chosen, particularly *Promax* rotation. This rotation method is recommended as a desirable oblique rotation choice [57,60] and is designed to result in a simple structure by producing greater correlations among factors [58].

Twenty-two items remained in the dataset following the removal of items due to low corrected item-total correlations and SMC, and each of the original eight CLASS factors (see Table II) was represented in the reduced dataset by at least two items. To investigate whether a similar factor structure could be obtained from our reduced dataset, our initial EFA sought to extract eight factors, and the resulting calculated communalities were examined. Communality is the sum of the squared correlations of the variable with the factors, and it represents the variance in the observed variables which are accounted for by a common factor [58]. Items with low calculated communalities do not load strongly onto any factor, and in this case one item was removed for having communality below 0.3. An eight-factor EFA was performed with the remaining 21 items, with all items possessing resulting calculated communalities above 0.3.

2. Factor retention and interpretation

In the first stage of factor extraction interpretation, the eigenvalues of the extracted factor loadings matrix were examined. Eigenvalues are the sum of squared factor loadings for a given factor, and multiple rules exist for eigenvalue cutoffs for factor retention. While the Kaiser criterion [57,61] suggests retaining factors with eigenvalues above 1.0, the Joliffe criterion [58,62] recommends a cutoff of 0.7. Each criterion would inform the retention of different factor quantities in this case, so a scree plot test with parallel analysis was performed. A scree plot is a graphical representation of the extracted eigenvalues for all factors, while a parallel analysis involves generating random uncorrelated data of the same size as the observed data sample, from which eigenvalues are also extracted. When comparing the eigenvalue scree plots from the observed and randomly generated datasets, the quantity of observed eigenvalues with values above the 95th percentile of the random eigenvalues should be retained. Results from the parallel analysis in this case supported the retention of two to four factors, depending on which criterion for eigenvalue cutoff is chosen—for instance, the Kaiser criterion suggested retaining only two factors, while the comparison between observed and random eigenvalues suggested retaining four (see Fig. 1). Our initial EFAs therefore sought to extract four factors.

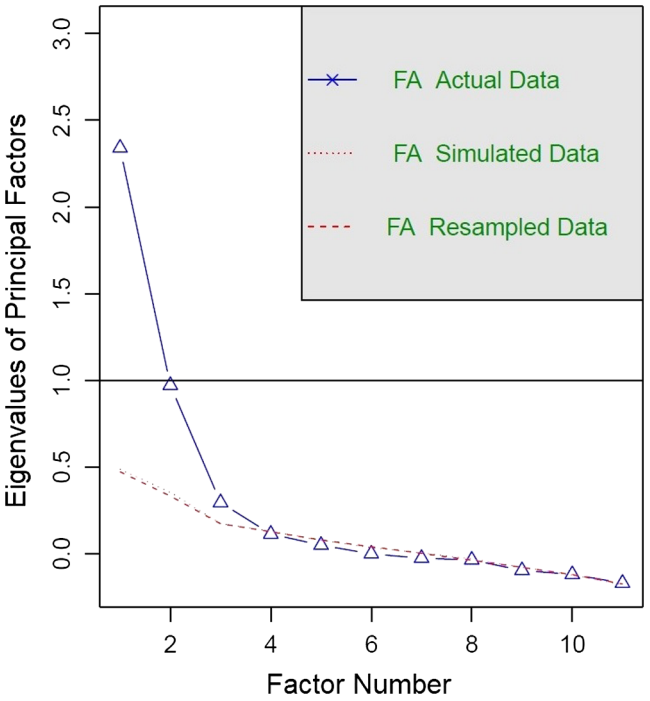

FIG. 1. Parallel analysis scree plot of 21-item survey, with overlaid Kaiser criterion cutoff; factor analysis (FA).

Examination of the factor loadings matrix for the eight-factor EFA indicated that there were items that loaded weakly (<0.3) onto multiple factors, three factors that had three or less items with significant loadings, and one item that only loaded significantly onto one factor. This initial output indicates that an eight-factor solution is not adequate, which aligns with the results of the parallel analysis shown in Fig. 1. The item that loaded onto an isolated factor was removed from the dataset, and a four-factor EFA was performed with the remaining 20 items. The factor loadings matrix from the 20-item, four-factor EFA

TABLE VI. Factor loadings from 20-item, four-factor EFA, sorted by factor. Loadings with magnitudes below 0.3 are omitted for clarity.

CLASS item Number	Factor 1	Factor 2	Factor 3	Factor 4
11	0.541	⋯	⋯	⋯
16	0.418	⋯	⋯	⋯
23	0.403	⋯	0.438	⋯
24	0.632	⋯	⋯	⋯
28	0.406	0.310	⋯	⋯
30	0.347	0.418	⋯	⋯
32	0.425	⋯	⋯	⋯
42	0.357	⋯	⋯	⋯
3	⋯	0.733	⋯	⋯
14	⋯	0.573	⋯	⋯
25	⋯	0.452	⋯	⋯
35	⋯	0.381	⋯	−0.307
37	⋯	0.365	⋯	⋯
5	⋯	⋯	0.692	⋯
13	⋯	⋯	0.388	⋯
29	⋯	⋯	0.447	⋯
40	⋯	⋯	0.563	⋯
15	⋯	⋯	⋯	0.477
34	⋯	⋯	⋯	0.455
36	⋯	⋯	⋯	0.488

can be seen in Table VI, where factor loadings are <0.3 omitted for clarity.

Adequacy test indices for the 20-item, four-factor EFA were within acceptable bounds. The root-mean square residual (RMSR) measures the standard deviation of model prediction errors, and a good fitting model would generate a narrow distribution of such errors. Thus, the calculated RMSM for a good fitting model would be close to zero—in our case it was 0.03. The root mean square error of approximation (RMSEA), which should be below 0.05, was 0.032; and the non-normed fit index, which analyzes the discrepancy between the chi-square value of the hypothesized model fit vs the chi-square value of a null model, was 0.954, where values >0.95 indicate good model fit. Although some items cross load onto multiple factors, this factor loadings matrix indicates that there are three factors with at least five significantly loading items and one factor with only four items, one of which with a negative loading. The item statements were then analyzed to interpret the factor loadings. All items associated with factor 1 refer to student attitudes and beliefs surrounding personal effort or motivation to understand physics concepts and formulas. This factor was thus called *"sense making effort and motivation."* All items associated with factor 2 refer to how physics understanding relates to students' lived experiences and personal enjoyment, this factor is therefore called *"personal connection and real-world application."* The last two factors relate to student attitudes and beliefs toward problem solving, with factor 3 possessing only items worded to align with novice beliefs, and factor 4 possessing only items worded to align with expert beliefs. Cronbach's α was calculated as a measure of internal reliability for each separate factor (see Table VII).

TABLE VII. Cronbach's α and average inter-item correlation for each factor in the 20-item, four-factor structure.

Factor	Cronbach's α	Average interitem correlation
1	0.75	0.27
2	0.75	0.30
3	0.66	0.28
4	0.49	0.20

These internal reliability estimates indicate that factor 4 is problematic. Following these results, in combination with the obtained eigenvalues that showed little difference between the third and fourth factors, as well as evidence from the factor loadings matrix suggesting that factor 4 does not possess enough significantly loading items to constitute a stable factor, we decided to investigate whether different three-factor structures could be extracted from the data. This led to multiple three-factor EFA attempts on the 20-item dataset, and initial factor loadings matrices supported the removal of items 34 and 35 for not loading significantly onto any of the three proposed factors. Results from the subsequent three-factor, 18-item EFA indicated that each item loaded onto at least one factor and given that each factor was theoretically sensible, this became the first potential model for CFA testing.

Three more potential competing EFA models were generated from the 20-item dataset. As the fourth factor extracted from the above four-factor, 20-item EFA was composed almost solely of items aligned with "expert" perspectives toward problem solving, we investigated whether a three-factor structure could adequately model the data if all items that loaded solely onto this factor were removed. This decision led to removing items 15, 34, and 36 from the dataset prior to iterative three-factor EFAs. The purpose of removing these items was to investigate whether a third extracted factor from the reduced set of items could be composed only of items aligned with novice perspectives toward problem solving, which would provide a more theoretically focused factor called "*attitudes toward problem solving.*" A three-factor, 17-item EFA followed, the results of which led to the removal of item 35 from the dataset as it did not load significantly onto any factor. Results from a three-factor, 16-item EFA provided appropriate overall adequacy test indexes and factor loadings for all items, and this factor structure became the second model proposed for CFA testing.

The third potential model was generated from the decision to remove item 23 from the three-factor, 16-item factor structure. Item 23 reads, "*In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than*

TABLE VIII. Internal reliability estimates and adequacy test indices for EFA factor structures.

Model	Full scale α (average interitem correlation)	Number of items in factor 1 (α) [average interitem correlation]	Number of items in factor 2 (α) [average interitem correlation]	Number of items in factor 3 (α) [average interitem correlation]	RMSR	RMSEA with 90% confidence intervals	Tucker-Lewis Index
18-item	0.81 (0.20)	9 (0.76) [0.26]	5 (0.66) [0.28]	5 (0.68) [0.30]	0.04	0.030, 0.042, 0.051	0.919
16-item	0.80 (0.20)	8 (0.75) [0.27]	6 (0.72) [0.31]	5 (0.66) [0.28]	0.03	0.020, 0.035, 0.046	0.950
15-item	0.78 (0.20)	7 (0.74) [0.29]	6 (0.72) [0.31]	5 (0.63) [0.25]	0.03	0.020, 0.035, 0.048	0.950
11-item	0.73 (0.2)	6 (0.72) [0.3]	5 (0.66) [0.28]	⋯	0.04	0.026, 0.045, 0.061	0.933

going back through the problem." Although this item loaded onto factors that made theoretical sense with its content, its descriptive statistics indicate that the vast majority of students answered it in a way that aligns with expert perspectives in both pre- and postapplications of the CLASS. A lack of a response pattern shift for a given item is not concerning in and of itself, but in light of the context of PEER Physics, we believe this item's usefulness is questionable. PEER Physics' inquiry-based curricular approach emphasizes the development of physics conceptual understanding instead of the quantitatively oriented problem solving that characterizes most physics classrooms. While the curriculum suite does include problem-solving modules, they are optional within the overall PEER Physics curricular progression. Problems included in the core curricular progression relate to how experimental evidence supports or refutes claims about physical phenomena, and within these problems, applications of formulas serve to emphasize their conceptual origins instead of single correct quantitative results. Item 23's wording targets number sense—the ability to understand whether an obtained quantitative result is contextually sensible—in connection to the sense-making effort during quantitative problem solving, and this sets it apart from the other items within its proposed factor. All other items whose statements relate to problem solving are worded so that their statements can relate to the physics problems that PEER Physics students engage with. After removing item 23, a three-factor, 15-item EFA was conducted. All items loaded onto conceptually appropriate factors and adequacy test indexes were acceptable, so this factor structure became the third potential model for CFA testing.

Although the different three-factor EFA structures all had acceptable adequacy text indexes, some of their items double-loaded onto multiple factors, and unequal amounts of items loaded onto each factor. This created difficulties in terms of factor and score interpretation, which worked against our overall objective with this work: to propose a survey factor structure with simpler interpretation, to support the development of a student attitudes and beliefs survey with greater usability in the high school context. To propose an even more parsimonious factor structure that might accomplish this objective, a fourth EFA structure was developed by seeking to extract only two factors, with more stringent conditions for factor loadings. Through an iterative process, EFAs were conducted, and items were removed if they did not meet a 0.4 threshold for their factor loading. This process resulted in a factor structure with diminished internal consistency indexes for both the entire scale and each factor, which was expected given the smaller number of items. However, items loaded onto conceptually appropriate factors and adequacy test indices were comparable to those of the other proposed factor structures. Furthermore, all items loaded strongly onto only one factor and there was no double loading, which aligns more closely with the overall objective of this study. The two-factor, 11-item structure thus became the fourth and last potential model for CFA testing.

Scree plots generated through parallel analysis for the 18-, 16-, and 15-item EFA structures all supported the extraction of three factors, and the interpretation of factor groupings is conceptually identical across each proposed model. However, the scree plot generated for the 11-item EFA structure was more ambiguous, suggesting either two or three factors, depending on which criterion was used to justify the factor extraction decision. The Kaiser criterion would support the extraction of two factors in this case. Table VIII provides more information on the four factor structures generated by EFA. These tables include internal reliability estimates for each model scale and factors, as well as calculated adequacy test indexes for each model's proposed factor structure.

In summary, our EFA analysis of pre-CLASS responses resulted in four potential factor structures, whose scree plots each support the extraction of either two or three factors. We now turn to a description of a CFA analysis that tested these structures, alongside those proposed by Cahill *et al.* [35] and Douglas *et al.* [34], on post-CLASS responses.

C. Confirmatory factor analysis

We initially ran six confirmatory factor analysis (CFA) models on our post-CLASS response data—one utilizing the Douglas *et al.* [34] factor structure extracted from undergraduate student CLASS responses, one utilizing the Cahill *et al.* [35] factor structure extracted from and for usage in interactive physics course, and four others to test

the factor structures extracted by our EFA analyses. Our rationale for testing the factor structures extracted from pre-CLASS response data on post-CLASS data was to decrease the possibility that CFA fit results could capitalize on the same chance variation [63]. In addition, as a pre or postinstrument, the factor structure should hold over time even if students experience shifts in their attitudes and beliefs during a course. In other words, if the factor structure is adequate, response patterns for respondents may shift over time, but the correlations between response patterns for items that relate structurally should not.

For model identification, we used *Mplus 7.4*. As responses to the CLASS are ordinal and cannot be assumed to follow a normal distribution, we used the weighted least square mean and variance (WLSMV) estimator, which uses a probit link function to account for the estimation of polychoric correlations between categorical variables. To compare how well the six models fit the data, we employed three of the global fit indices suggested by Kline [63]: model χ^2, Stieger-Lind root mean square error of approximation (RMSEA), and Bentler comparative fit index (CFI). The model χ^2 fit index tests the assumption that the covariance matrix calculated for the specified model is different from the population's actual covariance matrix, but this statistic is highly sensitive to sample size and is routinely violated in CFA. More practical measures include RMSEA, which is a badness-of-fit test calculated using the chi-square model statistic, but also accounts for degrees of freedom and sample size [63]. Values below 0.06 are considered to indicate good fit. CFI is a goodness-of-fit test that compares the amount of departure from close fit for the proposed model to the null model [63]. Values above 0.95 are considered to indicate good fit [64]. *Mplus* also provides a weighted root mean square residual (WRMR) value, which is a residual-based, badness-of-fit index suggested for categorical data instead of the more commonly applied standardized root mean square residual. A recent simulation study by DiStefano *et al.* [65] suggested that values below 1.0 indicate good model fit, which was aligned with previous recommendations for WRMR [66].

To investigate local fit, we instructed *Mplus* to incorporate residual correlation matrices and suggested modification indices in the model output. Since the factor indicators are ordinal with five categories, *Mplus* output included polychoric correlations, and correlation residuals above 0.1 may have indicated local misfit [63]. In addition, suggested modification indices with values above 10 that were theoretically consistent with the content and purpose of the instrument's factors were considered for model respecification. Potential modifications included the addition of correlations between error covariances, which may indicate similar sources of measurement error.

Finally, *Mplus* reported standardized parameter estimates for factor loadings, correlations between factors, error correlations, and a statistical significance test at the $p = 0.05$ value. Although we were more concerned about the overall fit of the six proposed models instead of a specific parameter loading, parameter estimates were a useful tool for assessing whether the empirical factor structure of the proposed models were consistent with the theoretical constructs the instrument attempts to measure. Parameter estimates were also useful for identifying potential locations of model misspecification.

1. CFA results

Descriptive statistics.—The average postresponses for the 21 CLASS items used in the six models ranged from 2.6 to 4.1. Eleven items had a median of 4, and eight items had a median of 3, indicating an overall positive skew to response distributions. Standard deviations were all close to 1.0, with a range of 0.89 to 1.1. Only item 24 was found to have skew or kurtosis with a magnitude larger than 1.0, indicating that their response distributions were concentrated around certain values. The relatively high medians among most of the items in the dataset, in combination with standard deviations close to 1.0 for all items, suggests non-normality of the data, which is consistent with the nature of the CLASS items and supports our decision to use the WLSMV estimator.

Global model fit statistics.—For comparative purposes, we list the model fit statistics for the Cahill *et al.* [35] and Douglas *et al.* [34] item structures alongside our proposed EFA models in Table IX.

Presented global fit indices suggest that the 11-item factor structure extracted from pre-CLASS data fits the post-CLASS data better than the factor structures proposed by Cahill *et al.* [35], Douglas *et al.* [34], or the other EFA analyses we conducted. Although all the models' χ^2 statistic had p values below 0.05, the χ^2 magnitude for the 11-item factor structure was the lowest ($\chi^2 = 119.2$). In addition, the 11-item factor structure had the CFI value closest to the 0.95 cutoff (CFI = 0.94), the RMSEA value closest to the 0.06 cutoff (RMSEA = 0.065), and a WRMR below the 1.0 cutoff (WRMR = 0.951). Following these results, we conducted a subsequent analysis with the goal of improving the 11-item factor structure through localized indicators of model misspecification.

Potential misspecifications to 11-Item EFA model.—There were three residual correlations between items that had absolute values greater than 0.1, indicating possible correlation overestimations: item 3 with item 14, item 23 with item 28, and item 5 with item 37. *Mplus* also produced a modification index greater than 10 associated with the residual correlation between item 3 and item 14, which suggests that adding error variance correlations between these indicators would improve overall model fit.

In addition to error variance correlations, modification indices suggested adding one cross loading to the model:

TABLE IX. Comparison of global fit indices across models.

CFA structure	χ^2	χ^2 dof	χ^2 p value	RMSEA (95% CI)	CFI	WRMR
Cahill *et al.*	948.5	274	0.00	0.076 (0.071–0.082)	0.80	1.589
Douglas *et al.*	294.2	83	0.00	0.077 (0.068–0.087)	0.89	1.230
18-item EFA	421.6	132	0.00	0.072 (0.064–0.080)	0.90	1.258
16-item EFA	418.8	101	0.00	0.086 (0.078–0.095)	0.87	1.368
15-item EFA	357.1	87	0.00	0.076 (0.077–0.095)	0.88	1.294
11-item EFA	119.2	43	0.00	0.065 (0.051–0.079)	0.94	0.951

cross loading item 30, which reads "*Reasoning skills used to understand physics can be helpful to me in my everyday life*," on a factor we named "*problem-solving practices*." Conceptually, the perceived everyday usefulness of physics reasoning skills could be related to a student's approach to solving physics problems. However, given our goals of simplifying the CLASS for usability and measuring distinct factors, we decided not to follow this modification suggestion and further examined the recommendation to include error variance correlations.

Respecification of EFA model.—Kline [63] states that the respecification of CFA models should be driven by "substantive consideration" rather than purely empirical grounds (p. 310). To add a correlation between the error correlations of items 3 and 14, there should be evidence that items have similar sources of measurement error. Item 3 reads "*I think about the physics I experience in everyday life"* and item 14 reads *"I study physics to learn knowledge that will be useful in my life outside of school."* Both items attempt to probe whether students relate physics to their lived experiences and use similar phrasing besides "everyday" versus "outside of school" to probe for a factor we named *"personal connections to physics."* Therefore, it seemed reasonable to include an error variance correlation between these items.

The simple addition of an error variance correlation between items 3 and 14 improved model fit to the point where three practical tests of global fit all reached the cutoff values suggested for good fit. The RMSEA was below the 0.06 cutoff (RMSEA = 0.056), and the WRMR was well below the 1.0 cutoff (WRMR = 0.851). This iteration of the model also met the 0.95 CFI criterion (CFI = 0.954). A summary of the influence of the modification on model fit is shown in Table X.

After adding the described re-specifications, the χ^2 test was still statistically significant ($p = 0.00$), indicating that there was a difference between the correlation matrix predicted by the model and the population correlation matrix. However, the factor structure we tested had relatively sparse loadings and correlations compared to most structures used in CFA analysis, resulting in relatively high degrees of freedom for the model, making it difficult to reach a χ^2 value that is not statistically significant. In addition, only two correlation residuals out of a possible 55 had an absolute value above 0.1, which is below the number of residuals over 0.1 we would expect by chance at $p = 0.05$. There were also no further modification indices above 10 that would not result in a negative factor cross loading. Following these results, we decided to tentatively accept the model shown in Fig. 2. In addition, Table XI shows the 11 CLASS items in our model and their associated factors.

Conceptual interpretation of factors.—As shown in Table XI and Fig. 2, our analyses resulted in a proposed structure with two factors that we named *personal connections to physics* and *problem-solving practices*. Factor 1, *personal connections to physics* (referred to as *"personal"* in following sections) consists of six items that elicit whether students are making connections between disciplinary physics ideas and their experiences outside of formal education settings. Items are drawn from two categories originally proposed by Adams *et al.* [6]: *Real world connection* and *personal interest*. All items are worded so that agreement with them is considered to indicate expertlike personal connection to physics. For example, if a student agrees with item 28, *"Learning physics changes my ideas about how the world works,"* there is evidence to suggest the student is beginning to think about and use physics to better understand their interactions with, and the behavior of, the natural world, much like an expert physicist might.

Factor 2, *problem-solving practices* (referred to as *"problem"* in following sections) consists of five items that elicit student perceptions of the efficacy of different

TABLE X. Impact of modifications on global fit indices.

Model modification	χ^2	χ^2 dof	χ^2 p value	RMSEA	CFI	WRMR
11-item EFA	119.2	43	0.00	0.065	0.93	0.951
11-item EFA with error correlation	96.8	42	0.00	0.056	0.95	0.851

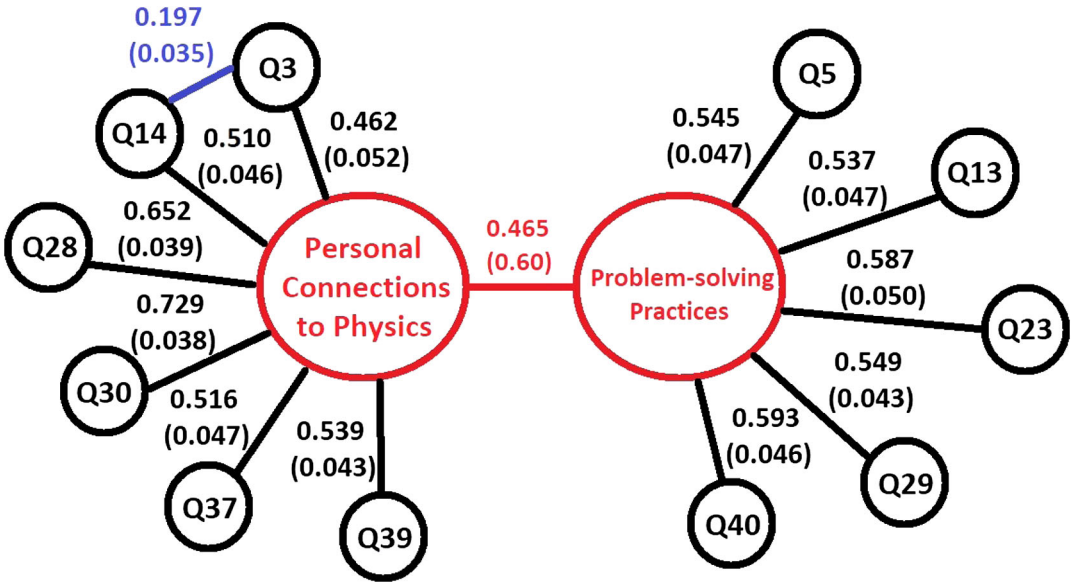

FIG. 2. Standardized factor loadings and correlations for the final model. Black numbers represent factor loadings; red numbers represent interfactor correlations; blue numbers represent interitem correlations; numbers in parentheses represent standard errors.

strategies for solving physics problems. We used the word practice when naming this factor to reflect that each item describes something that students actually or potentially do within a problem-solving context. Items 5, 13, and 40 come from three *problem-solving* categories proposed by Adams *et al.* [6], while item 23 was grouped with *"sense making or effort,"* and 29 was considered a stand-alone item. All items are worded so that disagreement with them is considered an expertlike perception of the efficacy of the described problem-solving practice. For instance, if a student disagrees with item 13, *"I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations,"* it may be an indicator of a student's expertlike beliefs about the usefulness of completing physics equations.

D. Replication analysis

Replication analyses serve to investigate whether a proposed factor structure for a given data set is likely to be observed within other datasets with similar characteristics [57]. The most basic threshold of replicability is replication of the basic factor structure, which can be ascertained by identifying whether item loadings remain congruent across data samples [57]. We performed an internal replication analysis to assess whether the 11-item,

TABLE XI. CLASS items in 11-item, two-factor model, sorted by factor.

Factor	CLASS items
Personal connection to physics	*3. I think about the physics I experience in everyday life.* *14. I study physics to learn knowledge that will be useful for my life outside of school.* *28. Learning physics changes my ideas about how the world works.* *30. Reasoning skills used to understand physics can be helpful to me in my everyday life.* *37. To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.* *39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem.*
Problem-solving practices	*5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.* *13. I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations.* *23. In doing a physics problem, if my calculation gives a result very different from what I expect, I'd trust the calculation rather than going back through the problem.* *29. To learn physics, I only need to memorize solutions to sample problems.* *40. If I get stuck on a physics problem, there is no chance I'll figure it out on my own.*

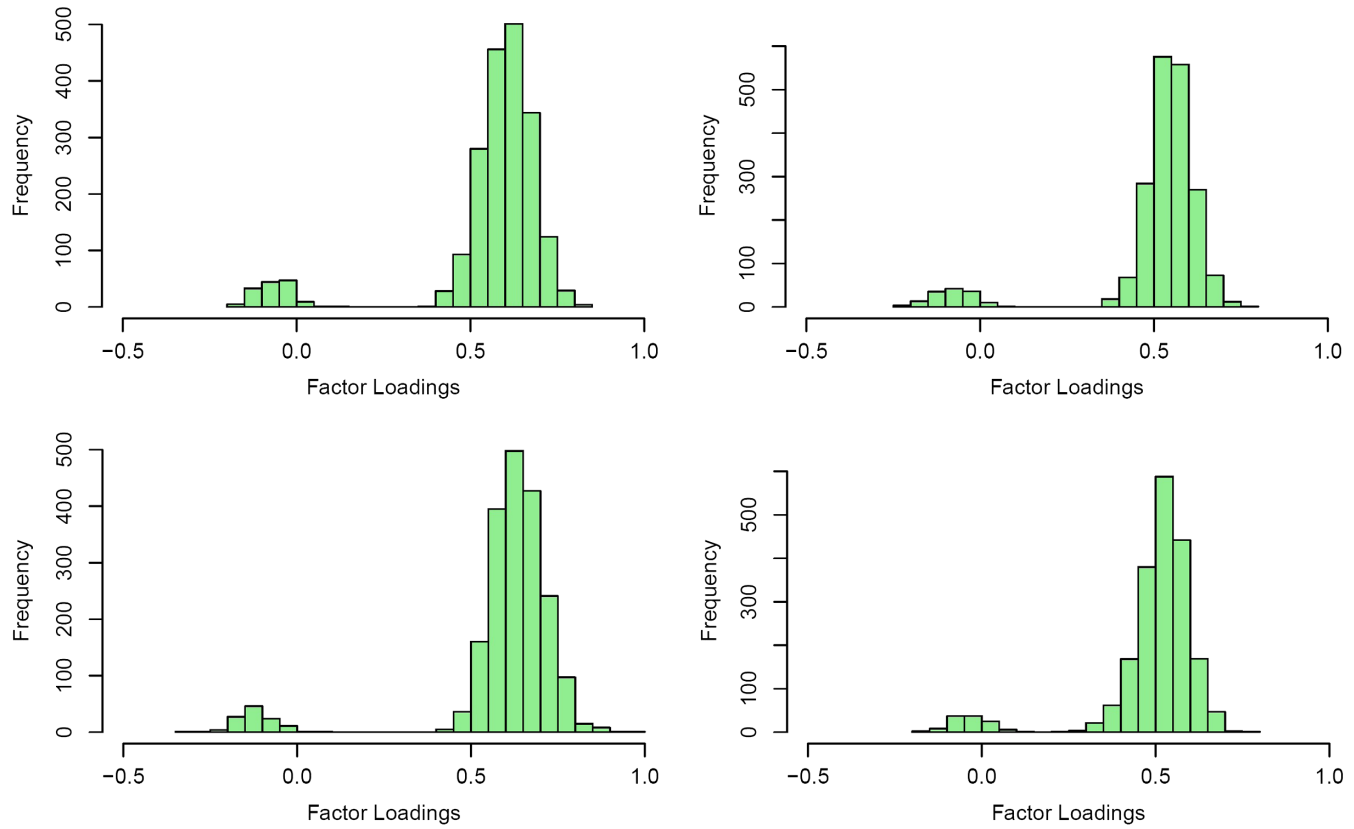

FIG. 3. Example distributions of factor loadings from replication analysis. Top row shows CLASS items 14 (left) and 37 (right) from factor 1 (personal), bottom row shows CLASS items 5 (left) and 29 (right) from factor 2 (problem).

two-factor model's basic factor structure would emerge within random subsamples of the full preresponse dataset. Subsamples of approximately equal size were obtained by randomly assigning respondents from the full dataset, and the same EFA procedure was applied on each subsample. Factor loadings were then compared to check if items loaded onto the same factors across subsamples.

EFA results are sensitive to sample size, and Osborne [57] recommends 20 cases for each variable present in an analysis as a minimum for generalizable results. There are 22 calculated factor loadings within our proposed 11-item, two-factor model, and thus an appropriate sample size to ascertain the replicability of such a model should contain at least 440 respondents. This is approximately the size of our full dataset, so it could be expected that factor loadings would be volatile when splitting the dataset further. Therefore, performing only one internal replication analysis could provide potentially untrustworthy results. To avoid this, we performed 2000 subsample EFA procedures and generated factor loading distributions for each item. We found that all items had factor loading distributions with one clear peak, which, given the sample size sensitivity of EFA, is an encouraging indication of the reduced survey model's potential stability. Figure 3 shows example factor loading histograms for two items in both factors—all items had nonvolatile behavior across subsamples.

E. Bootstrap resampling analysis

Given our relatively small dataset, bootstrap analyses offer a more appropriate path toward investigating the replicability of our potential factor structure. The resampling approach is designed for studies with inadequate samples [57], and it involves using an existing dataset to generate a certain number of related samples by randomly selecting and replacing subjects in the dataset. Unlike internal replication analysis, which involves splitting an existing dataset into multiple possible subsamples, bootstrapping resampling analysis generates new datasets by randomly selecting individual respondents from the master sample to be copied multiple times (or excised entirely). Thus, one can generate multiple samples of equal size, with slight variations between them, that allow the estimation of sampling distributions for various relevant statistics [60]. Generated resamples are all related since they derive from the same master sample but are varied in the sense that each individual from the master sample might be present in varying degrees (or not at all) in each. While resampling analyses cannot compensate for inappropriately small samples and can also promulgate biases that exist within a given dataset, they are able to provide confidence intervals and other forms of evidence toward the precision of a statistical solution [57].

TABLE XII. Bootstrapped coefficients and confidence intervals generated from the two-factor, 11-item EFA model. Bolded cells highlight item-factor pairings suggested by our proposed structure.

CLASS item Number	Factor 1 (personal) loading (confidence intervals)	Factor 2 (problem) loading (confidence intervals)
3	**0.56 (0.41, 0.70)**	−0.08 (−0.22, 0.08)
5	−0.13 (−0.24, −0.01)	**0.64 (0.50, 0.77)**
13	0.09 (−0.03, 0.21)	**0.41 (0.27, 0.55)**
14	**0.61 (0.47, 0.74)**	−0.08 (−0.27, 0.05)
23	0.02 (−0.09, 0.15)	**0.55 (0.39, 0.70)**
28	**0.57 (0.47, 0.68)**	0.06 (−0.05, 0.19)
29	−0.02 (−0.13, 0.11)	**0.53 (0.39, 0.66)**
30	**0.56 (0.45, 0.68)**	0.15 (0.04, 0.26)
37	**0.55 (0.43, 0.66)**	0.15 (0.04, 0.26)
39	**0.42 (0.31, 0.54)**	−0.08 (−0.19, 0.03)
40	0.02 (−0.09, 0.15)	**0.53 (0.40, 0.67)**

In this case, bootstrap resampling analyses allow us to investigate the precision of our obtained factor loadings, to investigate whether we can expect other, similar samples to produce similar results. For our study, we performed 5000 bootstrap replications of the two-factor, 11-item EFA on the full preresponse dataset and analyzed the confidence intervals of the calculated factor loadings, which are shown in Table XII.

Results from the bootstrap resampling analysis indicate that only one of the calculated factor loadings has a confidence interval where the lower bound is below the 0.30 threshold (item 13), which is evidence toward the replicability of the proposed factor structure. Given that all items have their upper confidence bound above 0.50, and that these results were obtained from the application of a much longer survey instrument, we believe that this factor structure is worth testing with high school student populations, to investigate the structural replicability of our proposed model. It is possible that conceptual connections between these items will become more statistically salient if they are implemented on their own in the context of a more parsimonious instrument, where other biasing effects such as survey fatigue may be mitigated.

VI. DISCUSSION: POTENTIAL USES OF THE SCALE

Our proposed 11-item survey structure contains items from five of the eight factors within the original CLASS. Of the six items comprising our *personal connections to physics* factor, four were originally from the *personal interest* CLASS factor (original CLASS items 3, 14, 28, and 30), one from the *real world connection* CLASS factor (original CLASS item 37), and one from the *sense making or effort* CLASS factor (original CLASS item 39). Of the five items comprising our *problem-solving practices* factor, two were originally from the *conceptual understanding* CLASS factor (original CLASS items 5 and 13), one from the *sense making or effort* CLASS factor (original CLASS item 23), one from the *applied conceptual understanding* CLASS factor (original CLASS item 40), and one that did not have a CLASS factor categorization (original CLASS item 29). Only the *problem solving (general)* and *problem solving (sophistication)* CLASS factors are not represented in our proposed survey instrument. It may be the case that the items in these factors probe aspects of physics problem solving that were not appropriate for the context of students in PEER Physics courses.

This study's objective is to describe the extraction of a parsimonious factor structure from a longer survey instrument's response data, and any discussion of real-world interpretations and analyses of student response patterns would require implementing the shorter instrument itself with the appropriate student population and collecting new data. However, we believe that it could be of value to provide examples, using our current data, of how student responses obtained with the hypothetical parsimonious instrument could be interpreted and used by researchers and practitioners. Some of these examples are discussed in the following subsections, and considering that these data are a subset of a greater dataset collected via a different, longer instrument, we emphasize that the empirical results below are not intended to provide evidence toward any rigorous argument regarding PEER Physics, or the attitudinal impacts of inquiry-based physics instruction.

A. Overall attitudinal shifts

The 11-item instrument has acceptable internal reliability ($\alpha = 0.73$) for a scale seeking to measure a single overall latent variable (e.g., attitudes and beliefs toward physics), and each of its items has acceptable average interitem correlation with each other item within the scale (*average interitem correlation* $= 0.2$). These two qualities lend themselves toward using students' "performance" on the 11-item instrument as a primary outcome of interest. If scored on a 5-point scale, the 11-item instrument has an outcome space with a maximum score of 55, which could be interpreted as "complete agreement with expert views of physics." In this investigation, we first scored the pre- and postapplications of the 11-item subset within our full sample using a 5-point scale; these results are shown in Table XIII and Fig. 4.

TABLE XIII. Descriptive statistics of scored pre- and postresponses to proposed 15-item instrument.

	Mean %	SD %	Median %	Min %	Max %	SE %
Pre	67.87	10.21	67.27	32.73	96.36	0.50
Post	68.83	11.06	69.09	30.91	96.36	0.54

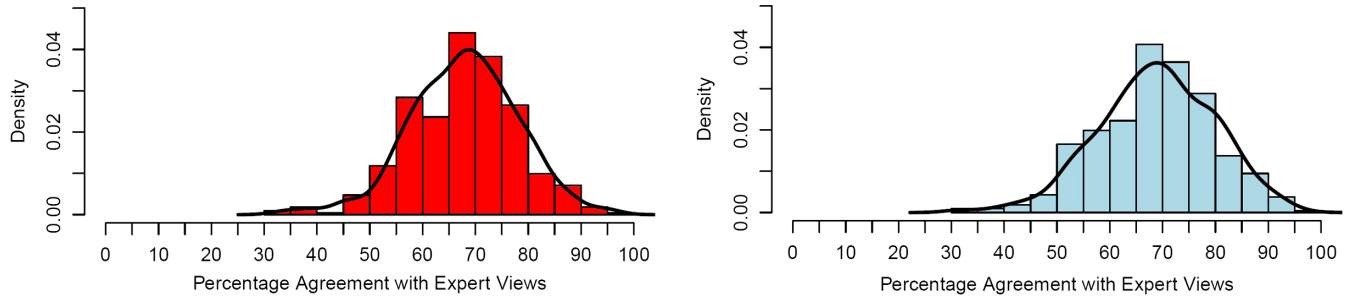

FIG. 4. Score distributions of pre (left) and post (right) responses to proposed 11-item instrument—bolded line represents the density curve for each histogram.

Score distributions of the 11-item subset instrument show very little overall change in students' attitudes and beliefs toward physics after an academic year of instruction. There are slight shifts in mean, minimum, and maximum scores between pre- to postapplications, but the overall score distributions are nearly identical. The distribution of score shifts (difference between post- and prescores) for the 11-item subset instrument is centered around 1%, with most shifts falling in the −150% to 150% range. These results can be seen in Table XIV and Fig. 5. To ascertain whether this small shift was statistically significant, a Shapiro-Wilk test of normality was performed on the pre- and postresponse distributions, which indicated that they were not normally distributed. Therefore, we performed a Wilcoxon paired-sample test on the distributions, whose results suggest the shift was indeed significant ($p < 0.001$).

Although prior research on the attitudinal impacts of physics instruction using the full CLASS instrument has indicated that inquiry-based instruction can be successful at fostering positive shifts in student expertlike beliefs [14], the results shown here indicate little change in such beliefs for high school students. However, this same meta-analysis [14] indicated that lecture-based courses tended to foster negative shifts in student expertlike beliefs, and in light of such results, no shifts at all (or very small positive shifts) in expertlike beliefs is encouraging, especially considering a high school population. This suggests that PEER Physics inquiry-based instruction did not generally impact high school students' attitudes and beliefs toward physics in a negative way as is usually the case in lecture-based courses.

These results indicate that following their course experience, there was a somewhat even split between students with positive and negative shifts in their attitudes and beliefs toward physics. Given the high school context, not all students necessarily intend to enroll in a university after their high school experience, and it may also be the case that for some of them, this was the first physics course they had ever taken. In this sense, even though these score distributions indicate a relatively small (or even nonexistent) overall shift, we believe these results to be positive given the school contexts and student demographics from which these data were collected. PEER Physics classrooms have an above-average proportion of students from historically underrepresented groups in physics. In addition, physics tends to be a mentally and emotionally challenging subject matter for most students. Therefore, results showing both negative and positive shifts in scores obtained from the 11-item instrument are not altogether surprising and using such results to try and understand why some students were positively impacted could be of worth for practitioners. We believe that such an instrument could provide instructors with data about aspects of students' experiences and internal motivation that traditional performance metrics such as quizzes and assessments do not seek to measure, and it could help highlight students who struggle conceptually, but otherwise possess the attitudes and beliefs that, if fostered, could encourage advancement in a scientific field.

TABLE XIV. Descriptive statistics for % shifts between pre- and postresponses to proposed 11-item instrument.

	Mean %	SD %	Median %	Min %	Max %	SE %
11-item instrument	0.96	11.40	1.82	−38.18	60.00	0.55

B. Instrument correlation with other relevant outcomes

Teachers who participated in data collection implemented the CLASS alongside the PEER Physics diagnostic

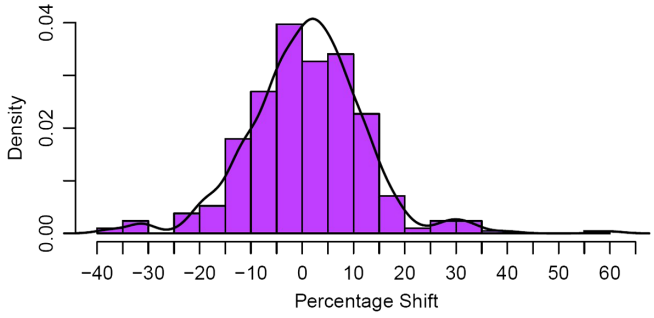

FIG. 5. Distributions of score shifts between pre- and postresponses of proposed 11-item instrument—bolded line represents the histogram density curve.

assessment, a conceptual exam which possesses a short survey component. One of the questions contained in the survey asks, "*Rate how much you enjoy learning science*." Responses to the survey questions are scored on a 5-point scale, and their results are used by some instructors to inform student groupings. The Spearman correlation, which indicates whether two variables are monotonically related, between students' responses to this survey question and their scores on the 11-item subset instrument were 0.33 in preresponses and 0.46 in postresponses. This indicates a positive, albeit weak, monotonic correlation between students' affect toward learning science and their attitudes and beliefs toward physics, a correlation which seemed to be strengthened after an academic year. This correlation provides first-order evidence of the construct validity of our hypothetical 11-item instrument, as it would be theoretically expected that students who self-report greater affect toward learning science would also showcase more expertlike thinking toward the connection science has to their own lives, as well as toward their own personal motivation to engage in it as an academic discipline.

C. Factor scores

As the 11-item instrument is composed of two factors, one could also investigate students' scores for each factor separately. Here we are calculating factor scores by summing scores from each item in the factor, and it should be noted that results from such an approach should be viewed with skepticism, as factor scores calculated in this manner are implicitly assuming that each item contributes equally toward the latent construct being measured by that factor [57]. EFA results themselves indicate that this assumption is difficult to hold, as some items have stronger factor loadings than others and thus should contribute more heavily to a hypothetical factor score. However, we have taken this approach here because this is how we believe instructors might use the scale in practice, and because calculating weighted factor scores in this context would be an unnecessarily complex approach given the exploratory nature of this study. Summed factor scores and score shifts between pre- and postapplications of the 11-item subset are shown in Table XV.

Factor 1 ($\alpha = 0.72$, average interitem $correlation = 0.3$), theorized as *personal connections to physics*, shows a slight increase in mean score between pre- and postapplications, while the median was unaffected, and the minimum score shifted downwards. Pre- and postresponse distributions for factor 1 were not normal under a Shapiro-Wilk test, while a Wilcoxon paired-sample test indicated that this shift was not statistically significant ($p = 0.559$). Factor 2 ($\alpha = 0.66$, average interitem $correlation = 0.28$), theorized as *problem-solving practices*, also shows a slight increase in mean scores between pre- and postapplications while the median was unaffected. However, in factor 2 both the minimum and maximum shifts were more pronounced, indicating more dramatic impacts on students' attitudes toward problem solving. Pre- and postresponse distributions for factor 2 were also not normal under a Shapiro-Wilk test, however, a Wilcoxon paired-sample test indicated that this shift was statistically significant ($p < 0.05$). We find the slightly larger mean positive shift in factor 2 somewhat unexpected, albeit encouraging, given that problem solving is an aspect of science practice that tends to be negatively viewed by students. Factor score distributions and shifts can be seen in Figs. 6 and 7.

TABLE XV. Descriptive statistics of personal and problem factor scores and shifts from pre- and postresponses to proposed 11-item instrument.

	Mean %	SD %	Median %	Min %	Max %	SE %
Personal–Pre	65.24	12.24	66.67	33.33	96.67	0.60
Personal–Post	65.63	14.02	66.67	23.33	96.67	0.68
Personal–Shift	0.39	14.92	0	−46.67	50.00	0.73
Problem–Pre	71.03	13.34	72.00	24	100.00	0.65
Problem–Post	72.67	12.89	72.00	28.00	100.00	0.63
Problem–Shift	1.65	13.63	0	−48.00	72.00	0.66

Student *personal* factor scores and shift distributions for both pre- and postscores suggest that overall, students were not impacted with regards to this factor. However, the distribution of factor score shifts tells a different story, one in which there were large shifts in individual students' scores related to this factor, a story that is hidden by the distribution's approximately normal shape. This is an example of the importance of obtaining matched sample data when implementing attitudes and beliefs surveys, as this allows practitioners and researchers to see individual effects that might be obfuscated under a population perspective. In this case, we see that there were students with negative shifts of up to −45% and positive shifts of up to 50%, dramatic results that the first two plots on their own would not indicate.

The distributions of student factor scores and shifts for *problem-solving practices* indicate, somewhat surprisingly, that student attitudes and beliefs toward problem solving in this population were positive to begin with and tended to stay that way overall. This might be due to response bias, as students could be hesitant to respond negatively to statements that could reflect on them poorly in an academic sense. As was the case with *personal connections to physics*, this factor had dramatic positive and negative shifts that the pre- and postscore distributions alone would not have indicated. Having the opportunity to identify individual students with such shifts may be of value to instructors, especially considering the negative views that students tend to have of problem solving.

VII. LIMITATIONS

This study's main limitations stem from the population sample and the exploratory and sample-size dependent

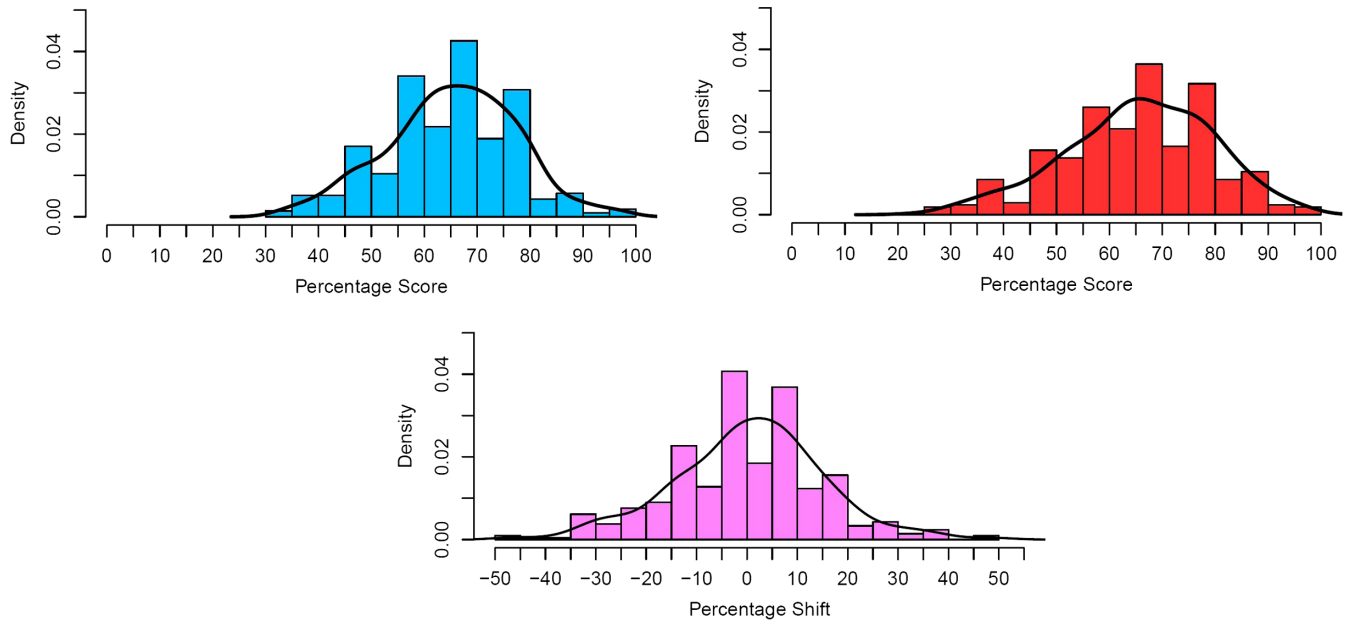

FIG. 6. Distributions of factor scores and shifts for *personal connections to physics*. Top row shows score distributions for pre (left) and post (right) responses, bottom row shows the distribution of score shifts. The bolded line represents the density curve for each histogram.

nature of the EFA procedures used to propose the tested factor structures. Decisions made during data cleaning reduced the sample size significantly and might have generated significant selection bias, as students who responded to all PEER Physics survey and CLASS items might also tend to be more academically oriented in general, or already have above-average positive attitudes and beliefs toward science and physics. However, these decisions seemed sensible given the sample population and optional nature of the original CLASS implementation. Instructors partnered with PEER Physics teach at schools serving a wide variety of student demographics and socioeconomic status, and all student data was combined for the purposes of the study.

We strongly emphasize the "exploratory" aspect of EFA, as our study does not intend to propose the 11-item,

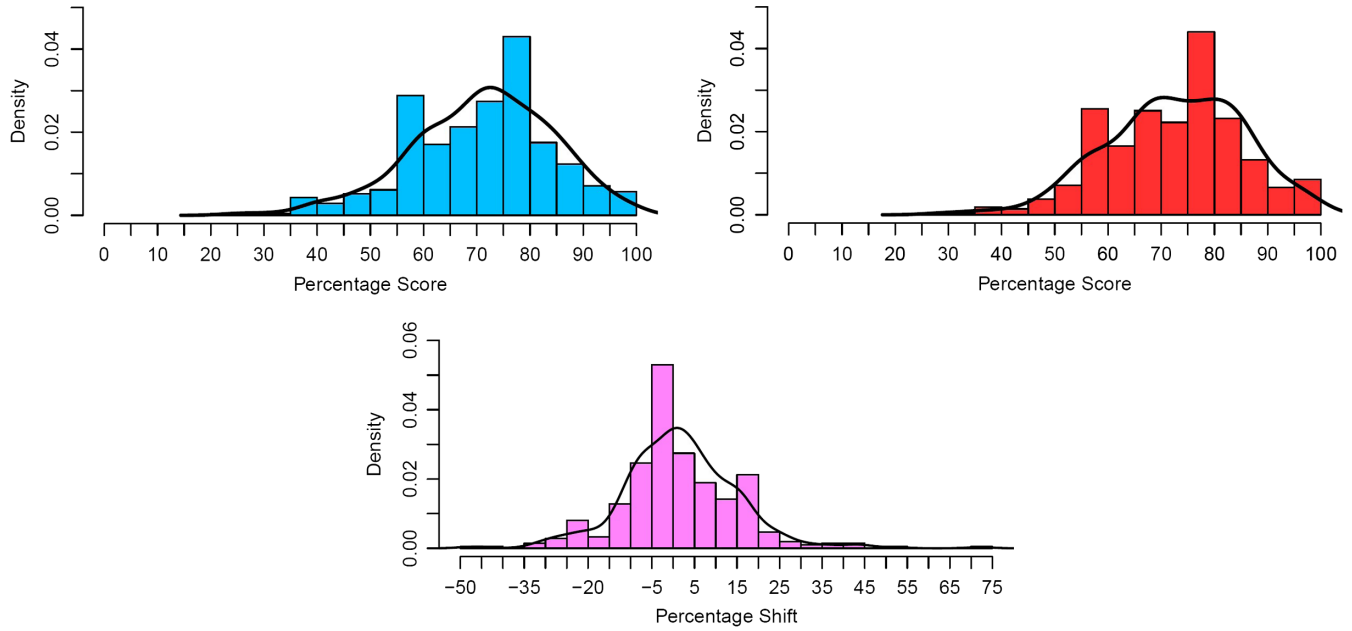

FIG. 7. Distributions of factor scores and shifts for *problem-solving practices*. Top row shows score distributions for pre (left) and post (right) responses, bottom row shows the distribution of score shifts. The bolded line represents the density curve for each histogram.

two-factor structure as an improved version of the CLASS for all purposes. Rather, we contend that this more parsimonious survey structure seems appropriate for the curricular approach and student population of PEER Physics, as well as for other similarly NGSS-aligned classroom environments that focus on engaging students in science practices. This study intends to support future empirical research regarding student attitudes and beliefs toward physics in these contexts. Researchers interested in taking inspiration from the CLASS to collect attitudinal data from students in different contexts should undergo similar data analyses and procedures as we have described. Different student populations might interact differently with the survey items, which implies that different item structures could be more appropriate in those contexts. This is exemplified by how our results differ from those of Douglas *et al.* [34] and Cahill *et al.* [35], whose studies were conducted with markedly different student populations.

VIII. CONCLUSION AND NEXT STEPS

The goal of this study was to ascertain the psychometric properties of high school student responses to the CLASS, and to use these properties to propose a more practical survey factor structure for measuring high school students' attitudes and beliefs about physics. Our proposed factor structure is more parsimonious than the original CLASS and fits a sample of high school responses better than the model proposed by Douglas *et al.* [34] and Cahill *et al.* [35] from samples of undergraduate students. This might be expected given the pronounced demographic differences between our student populations. Where Douglas *et al.*'s [34] sample was disproportionately white and male, ours has a relatively high proportion of students who self-identified as Hispanic, as well as a more even split between genders. Cahill *et al.*'s [35] sample is more evenly split in terms of gender, but with a smaller proportion of students underrepresented in physics than our own. In turn, this factor structure may provide initial validity evidence for a shorter measure of high school students' attitudes and beliefs toward physics.

From a practical perspective, our proposed instrument may be easier for instructors to implement, score, and interpret than the original CLASS. This is important for high school teachers with busy schedules, large teaching loads, and limited capacity for data use. Teachers could use the information gleaned from this measure to adjust their pedagogy to account for students' attitudes and beliefs toward physics, as traditional performance metrics do not target this aspect of the student experience. From a psychometric perspective, this instrument may also have more easily interpretable properties for usage with high school student populations than the original CLASS.

Testing these hypotheses would require implementing our proposed survey instrument with different samples of high school students, some ideally involving students in non-PEER Physics contexts, in order to see if the factor structure proposed in this paper holds across similar groups. Once the model structure is confirmed or respecified, it would also be beneficial to measure its usage and impact across a large sample of high school teachers, so that the instrument's value as a practical measure can also be ascertained. We anticipate that high school teachers will find this instrument to be more practically useful than the original CLASS, as it has increased utility to inform both day-to-day pedagogical decisions and provide evidence for longitudinal comparisons of long-term instructional improvement efforts. Supporting this hypothesis is the relative ease of usage of a 11-item survey with two clear factors in comparison to a 42-item survey with a complex factor structure, alongside the validity evidence presented in this paper. Future validation work or studies involving applications of the CLASS with high school populations should also seek to investigate how high school students interpret the item statements themselves, similarly to how Sawtelle, Brewe, and Kramer [67] did for undergraduate students at a primarily Hispanic-serving institution, as there may be common misunderstandings that affect their responses to statements and thus skew survey results. While the lack of more than one discernible maxima in the item response distributions generated by our replication analysis suggests that the students in our sample were interpreting items similarly, whether that interpretation is akin to what the developers of the CLASS intended should be specifically investigated.

Prior research involving the CLASS has argued for the importance and potential long-term impacts of the affective response of students both to physics instruction, and to the perceived significance that physics has to their own lives [31]. We agree that attending to high school students' attitudes and beliefs toward physics is an important step toward improving the physics instruction found in high school classrooms. This research contributes to this goal by providing first-order evidence that the factor structure of a more parsimonious version of the CLASS fits response patterns provided by a high school student population, which might draw greater scholarly attention to the potential value of collecting data on the attitudes and beliefs of students prior to the university context.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Valerie Otero for her support and guidance during the development of this manuscript. They also acknowledge Dr. Erin Furtak for providing the space in which this project was initiated, the teachers partnered with PEER Physics and their students for their willingness to collect and share data, as well as NSF-DUE Grants No. 1525338 and No. 1557351 for partially funding this work.

[1] V. K. Otero and D. E. Meltzer, 100 years of attempts to transform physics education, Phys. Teach. **54**, 523 (2016).

[2] D. E. Meltzer and V. K. Otero, A brief history of physics education in the United States, Am. J. Phys. **83**, 447 (2015).

[3] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[4] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).

[5] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[6] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. Phys. Educ. Res. **2**, 010101 (2006).

[7] https://www.physport.org/assessments/assessment.cfm?A=CLASS.

[8] K. Semsar, J. K. Knight, G. Birol, M. K. Smith, and D. K. O'Dowd, The Colorado learning attitudes about science survey (CLASS) for use in biology, CBE Life Sci. Educ. **10**, 268 (2011).

[9] 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, J. Chem. Educ. **85**, 1435 (2008).

[10] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, Development and validation of the Colorado learning attitudes about science survey for experimental physics, AIP Conf. Proc. **1513**, 442 (2012).

[11] S. J. Pollock, No single cause: Learning gains, student attitudes, and the impacts of multiple effective reforms, AIP Conf. Proc. **790**, 137 (2005).

[12] V. K. Otero and K. E. Gray, Learning to think like scientists with the PET curriculum, AIP Conf. Proc. **951**, 160 (2007).

[13] E. Brewe, L. Kramer, and G. O'Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS, Phys. Rev. Phys. Educ. Res. **5**, 013102 (2009).

[14] A. Madsen, S. B. McKagan, and E. C. Sayre, How physics instruction impacts students' beliefs about learning physics: A meta-analysis of 24 studies, Phys. Rev. Phys. Educ. Res. **11**, 010115 (2015).

[15] H. Alhadlaq *et al.*, Measuring students' beliefs about Physics in Saudi Arabia, AIP Conf. Proc. **1179**, 69 (2009).

[16] L. Ding, A comparative study of middle school and high school students' views about physics and learning physics, AIP Conf. Proc. **1513**, 118 (2012).

[17] K. A. Slaughter, S. P. Bates, and R. K. Galloway, A longitudinal study of the development of attitudes and beliefs towards physics, AIP Conf. Proc. **1413**, 359 (2011).

[18] P. Zhang and L. Ding, Large-scale survey of Chinese precollege students' epistemological beliefs about physics: A progression or a regression?, Phys. Rev. Phys. Educ. Res. **9**, 010110 (2013).

[19] K. A. Slaughter, S. P. Bates, and R. K. Galloway, How attitudes and beliefs about physics change from high school to faculty, Phys. Rev. Phys. Educ. Res. **7**, 020114 (2011).

[20] V. K. Otero and K. E. Gray, Attitudinal gains across multiple universities using the Physics and Everyday Thinking curriculum, Phys. Rev. Phys. Educ. Res. **4**, 020104 (2008).

[21] https://www.physport.org/assessments/assessment.cfm?A=EBAPS.

[22] K. Heredia and J. E. Lewis, A psychometric evaluation of the Colorado learning attitudes about science survey for use in chemistry, J. Chem. Educ. **89**, 436 (2012).

[23] L. Crocker and J. Algina, *Introduction to Classical & Modern Test Theory* (Cengage Learning, Mason, OH, 2008).

[24] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. Phys. Educ. Res. **1**, 010101 (2005).

[25] K. K. Perkins *et al.*, Correlating student beliefs with student learning using the Colorado Learning Attitudes about Science Survey, AIP Conf. Proc. **790**, 61 (2005).

[26] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[27] K. E. Gray, W. K. Adams, C. E. Wieman, and K. K. Perkins, Students know what physicists believe, but they don't agree: A study using the CLASS survey, Phys. Rev. Phys. Educ. Res. **4**, 020106 (2008).

[28] J. de la Garza and H. Alarcon, Assessing students' attitudes in a college physics course in Mexico, AIP Conf. Proc. **1289**, 129 (2010).

[29] B. Van Dusen and J. Nissen, Criteria for collapsing rating scale responses: A case study of the CLASS, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2019), pp. 585–590.

[30] American Physical Society, Diversity Statement, (2008, revised 2018). https://www.aps.org/policy/statements/08_2.cfm.

[31] K. K. Perkins and M. Gratny, Who becomes a physics major? A long-term longitudinal study examining the roles of pre-college beliefs about physics and learning physics, interest, and academic achievement, AIP Conf. Proc. **1289**, 253 (2010).

[32] F. Goldberg, V. Otero, and S. Robinson, Design principles for effective physics instruction: A case from physics and everyday thinking, Am. J. Phys. **78**, 1265 (2010).

[33] B. A. Lindsey, L. Hsu, H. Sadaghiani, J. W. Taylor, and K. Cummings, Positive attitudinal shifts with the physics by inquiry curriculum across multiple implementations, Phys. Rev. Phys. Educ. Res. **8**, 010102 (2012).

[34] K. A. Douglas, M. S. Yale, D. E. Bennett, M. P. Haugan, and L. A. Bryan, Evaluation of Colorado learning attitudes about science survey, Phys. Rev. Phys. Educ. Res. **10**, 020128 (2014).

[35] M. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. Phys. Educ. Res. **10**, 020101 (2014).

[36] C. E. Wieman and W. K. Adams, On the proper use of statistical analyses; a Comment on "Evaluation of Colorado Learning Attitudes about Science Survey" by Douglas et al, arXiv:1501.03257v1.

[37] E. B. Mandinach and E. S. Gummer, A systemic view of implementing data literacy in educator preparation, Educ. Res. **42**, 30 (2013).

[38] S. Cosner, Teacher learning, instructional considerations and principal communication: Lessons from a longitudinal study of collaborative data use by teachers, Educ. Manage. Admin. Leadership **39**, 568 (2011).

[39] E. N. Farley-Ripple and J. L. Buttram, Developing collaborative data use through professional learning communities: Early lessons from Delaware, Stud. Educ. Eval. **42**, 41 (2014).

[40] S. Blanc, J. Christman, R. Liu, C. Mitchell, E. Travers, and K. E. Bulkley, Learning to learn from data: Benchmarks and instructional communities, Peabody J. Educ. **85**, 205 (2010).

[41] T. J. Lasley, *Handbook of Data-Based Decision Making in Education* (Routledge, New York, 2009).

[42] B. Means *et al.*, Centre for Learning Technology Report, 2009, https://repository.alt.ac.uk/629/.

[43] M. I. Honig and N. Venkateswaran, School-central office relationships in evidence use: Understanding evidence use as a systems problem, Am. J. Educ. **118**, 199 (2012).

[44] A. S. Bryk *et al.*, *Learning to Improve: How America's Schools Can Get Better at Getting Better* (Harvard Education Press, Cambridge, MA, 2015).

[45] W. R. Penuel *et al.*, Developing a validity argument for practical measures of student experience in project-based science classrooms, in *Proceedings of the Annual Meeting of the American Educational Research Association, New York* (AERA, Washington, DC, 2018).

[46] W. R. Penuel and D. A. Watkins, Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education, Ann. Am. Acad. Political Social Sci. **683**, 201 (2019).

[47] M. Hannan, J. L. Russell, S. Takahashi, and S. Park, Using improvement science to better support beginning teachers: The case of the building a teaching effectiveness network, J. Teach. Educ. **66**, 494 (2015).

[48] National Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Academies Press, Washington, DC, 2012).

[49] NGSS Lead States, *Next Generation Science Standards: For States, By States* (National Academies Press, Washington, DC, 2013).

[50] S. N. Belleau and V. K. Otero, Critical classroom structures for empowering students to participate in science discourse, AIP Conf. Proc. **1513**, 11 (2012).

[51] M. Ross and V. K. Otero, Challenging traditional assumptions of secondary science through the PET curriculum, AIP Conf. Proc. **1513**, 350 (2012).

[52] S. N. Belleau and V. K. Otero, Scientific practices: Equalizing opportunities for linguistically diverse student groups, in *Proceedings of the 2013 Physics Education Research Conference, Portland, OR*, edited by P. V. Engelhardt, A. D. Churukian, and D. L. Jones (AIP, New York, 2013), pp. 67–72.

[53] W. E. Lindsay, V. K. Otero, and S. N. Belleau, PEER Suite: A holistic approach to supporting inductive pedagogy implementation, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC*, edited by A. Traxler, Y. Cao, and S. Wolf (2018), pp. 11–14.

[54] W. E. Lindsay, S. Widman, and M. Garcia, The association between sustained professional development and physics learning, in *Physics Education Research Conference Proceedings, Provo, 2019*, edited by Y. Cao, S. Wolf, and M. B. Bennett (AIP, New York, 2018), pp. 318–323.

[55] E. Cristobal, C. Flavián, and M. Guinalíu, , Perceived e-service quality (PeSQ): Measurement validation and effects on consumer satisfaction and web site loyalty, Managing Service Quality: An Int. J. **17**, 3 (2007).

[56] J. E. Francis and L. White, PIRQUAL: A scale for measuring customer expectations and perceptions of quality in internet retailing, in *Proceedings of the American Marketing Association's Winter Educators' Conference, Chicago, 2002*, edited by K. R. Evans and L. K. Scheer (American Marketing Association, Chicago, 2002), pp. 263–270.

[57] J. W. Osborne, *Best Practices in Exploratory Factor Analysis* (CreateSpace Independent Publishing Platform, 2014).

[58] A. G. Yong and S. Pearce, A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutorials Quant. Methods Psychol. **9**, 79 (2013).

[59] A. B. Costello and J. Osborne, Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis, Pract. Assess. Res. Eval. **10**, 7 (2005).

[60] B. Thompson, *Exploratory and Confirmatory Factor Analysis* (American Psychological Association, Washington, DC, 2004).

[61] H. F. Kaiser, A second generation little jiffy, Psychometrika **35**, 401 (1970).

[62] I. T. Joliffe, *Principal Component Analysis* (Springler Verlag, New York, 1986).

[63] R. Kline, *Principles and Practice of Structural Equation Modeling* (Guilford Press, New York, 2016).

[64] L. T. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Struct. Eq. Modeling: A Multidisc. J. **6**, 1 (1999).

[65] C. DiStefano, J. Liu, N. Jiang, and D. Shi, Examination of the weighted root mean square residual: Evidence for trustworthiness?, Struct. Eq. Modeling: A Multidisc. J. **25**, 3 (2018).

[66] C. Y. Yu, *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes* (University of California, Los Angeles, 2002).

[67] V. Sawtelle, E. Brewe, and L. Kramer, Validation study of the Colorado Learning Attitudes about Science Survey at a Hispanic-serving institution, Phys. Rev. Phys. Educ. Res. **5**, 023101 (2009).