Evaluation of Colorado Learning Attitudes about Science Survey

K. A. Douglas,[1] M. S. Yale,[2] D. E. Bennett,[3] M. P. Haugan,[4] and L. A. Bryan[4,5]

[1]*Engineering Education, Purdue University, West Lafayette, Indiana 47907, USA*
[2]*College of Business, University of Dallas, Irving, Texas 75062, USA*
[3]*Department of Educational Studies, Purdue University, West Lafayette, Indiana 47907, USA*
[4]*Department of Physics, Purdue University, West Lafayette, Indiana 47907, USA*
[5]*Department of Curriculum and Instruction, West Lafayette, Indiana 47907, USA*

(Received 20 September 2013; revised manuscript received 4 June 2014; published 19 November 2014; publisher error corrected 5 December 2014)

The Colorado Learning Attitudes about Science Survey (CLASS) is a widely used instrument designed to measure student attitudes toward physics and learning physics. Previous research revealed a fairly complex factor structure. In this study, exploratory and confirmatory factor analyses were conducted on data from an undergraduate introductory physics course ($n = 3844$) to determine whether a more parsimonious factor structure exists. Exploratory factor analysis results indicate that many of the items from the original CLASS have poor psychometric properties and could not be used in a revised factor structure. The cross validation showed acceptable fit statistics for a three factor model found in the exploratory factor analysis. This research suggests that a more optimum measurement of students' attitudes about physics and learning physics is obtained with a 15-item instrument, which describes the factors of personal application, personal effort, and problem solving. The proposed revised version of the CLASS offers researchers the opportunity to test a shortened version of the instrument that may be able to provide information about students' attitudes in the areas of personal application of physics, personal effort in a physics course, and approaches to problem solving.

DOI: 10.1103/PhysRevSTPER.10.020128 PACS numbers: 01.40.Fk, 01.40.G-, 01.40.Di, 01.50.Kw

I. INTRODUCTION

The advancement of knowledge in science education is dependent upon the quality of the assessments that are used in research. Pelligrino [1] has argued that there is a need in science education for rigorously developed and tested assessments that provide multiple sources of validity evidence. While the word "validated" is often used to describe assessments, validity more correctly refers to the extent to which theoretical and empirical evidence supports the interpretation of results when properly used [2]. Simply defined, validity means a test measures what it is intended to measure [3]. This is an important distinction because validity is inferred when multiple sources of evidence about different aspects of an instrument have been obtained. Validity is not a one-time stamp of approval made about an assessment, but rather is inferred based on the evidence [4]. According to Kubiszyn and Borich, there is "validity evidence if we can *demonstrate* that it [assessment] measures what it says it measures," ([3] p. 306).

Furthering the discussion on validity, Messick argued that validation is an ongoing pursuit for the purpose of improving and refining an instrument [5]. Based on Messick's discussion, assessments should be viewed through a developmental lens. Others have noted this and acknowledged that it may take years of research and several iterations before an assessment's psychometric properties become acceptable [6]. Pointed out by Arjoon, Xu, and Lewis, "If the instruments' scores are not valid and reliable, the resulting interpretations will also be invalid, and can lead to potential detrimental decisions for students and for the research or educational enterprise," ([7], p. 536). It is not uncommon for an assessment to be strong in one or more aspects of validity, and yet still need further development to adequately address the necessary aspects before the scores should be used for educational decisions or research. In other fields, such as engineering education and chemistry education, revised scales have been published in order to inform the community about psychometric issues and further develop and improve the assessment [6]. In the same way that replicating a scientific study provides evidence of the trustworthiness of the original results, replication studies of important assessment tools also enhance the degree to which researchers can be confident in their results obtained by the assessment. Once an instrument has been published for community use, it is the researchers' responsibility, not only the creators, but also those who use that instrument to examine validity evidence with their data through psychometric analysis. Considering that research results are dependent upon measurement, it is ethical to inform the community about any evidence that calls into question the validity of the

results for widely used assessments. The purpose of this study is to examine the psychometric properties of an important assessment in the research of students' attitudes, Colorado Learning Attitudes About Science Survey (CLASS) [8], and propose revisions based on evidence found. By first establishing which items robustly measure the same constructs and can therefore appropriately be scaled, this study aims to improve the validity of the interpretation and use of the CLASS results.

II. ASSESSMENT FUNDAMENTALS

The extent to which an assessment produces valid results is dependent upon the quality of the methods used in its development. Assessments that have been developed from rigorously applying methods should demonstrate evidence of validity. The methods of development can vary from qualitatively driven approaches (see Creswell and Plano Clark's discussion of exploratory sequential mixed methods [9]) to the more traditional, theoretically based approaches; see Spector [10]). What determines the quality of an assessment is not the methodological approach *per se*, but rather the rigor of the methods used. While there is a plethora of assessment development models that researchers can choose from to design their research methods, there are fundamentals that must be adhered to in order to assert something is in fact being measured by the items in an assessment.

The construct to be assessed must be sufficiently defined with a clear scope prior to item generation [4]. This can be done through qualitative approaches, such as interviews with students and experts to understand the variation that exists between them; or this can be done based on theoretical definitions. What is important is that the creators have a well-defined construct (or constructs) and have established the bounds of which the construct is to be studied [4]. In particular, latent constructs, such as attitudes, cannot be measured directly and, therefore, need several indicators in order to produce valid inferences [6]. Analogous to taking several measurements of a physical phenomenon in order to reduce measurement error, also in assessments, measurement error is reduced by increasing the number of measurements, i.e., questions about attitudes. However, when measuring attitudes, the actual phenomenon is not observable. Rather, what we have are questions that when taken together, theoretically represent some aspect of the phenomenon (e.g., attitudes) under study. The questions asked are a sample, or chosen subset, of all the questions that could be asked related to the phenomenon. The more fully all the aspects of the phenomenon are represented, the less the measurement error. In other words, researchers decide what questions need to be asked so that when taken all together, the questions will generalize to represent the phenomenon. This is a similar rationale as in physical study; researchers employ a strategy to ensure the results generalize to the phenomenon. Because the study of latent constructs, such as attitudes, is not directly measureable, care must be taken to examine the evidence that items labeled as a construct are in fact representative of the true phenomenon. By clearly defining the bounds of the phenomenon under study, researchers can articulate the purpose and scope of the assessment and go on to create items that are reflective of this purpose.

The development stage requires a clear rationale for why the chosen items are considered to be measures of the construct and then empirical evidence should demonstrate their adequacy in representing the construct. The first empirical evidence and prerequisite to justify the use of a scale is through examination of dimensionality [4]. According to Nunnally and Bernstein, the statistical properties of items that measure the same construct will indicate unidimensionality [11]. In addition, Netemeyer, Bearden, and Sharma wrote, "Given that scale (factor) unidimensionality is considered prerequisite to reliability and validity, assessment of unidimensionality should be paramount" ([4], p. 9). Statistically speaking, dimensionality is defined as the number of latent constructs (factors) needed to explain the item correlations [4]. Evidence of unidimensionality can be found through factor analysis, where constructs have a clear factor structure with no or minimal cross loadings [12]. A factor that is comprised of only items that cross load on other factors indicates an ill-defined construct. In addition, internal reliability will be high, as reliability is in part based on item variance [4].

Cronbach's alpha is one commonly used method of ascertaining the internal reliability of the assessment [13]. It provides evidence of the degree to which items are all measuring the same construct through the computation

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\alpha}_i^2}{\hat{\alpha}_X^2}\right),$$

where k is the number of items in the scale, $\hat{\alpha}_i^2$ is the variance of item i, and $\hat{\alpha}_X^2$ is the total variance of the scale [14]. While alpha is a function, in part, of the total number of items, it is also dependent on the variance attributable to persons and to the interaction between persons and items [13]. Cortina studied the relationship between coefficient alpha, dimensionality, the number of items in a scale, and average interitem correlations [15]. He found that for a unidimensional scale where the average interitem correlation is 0.50, alpha will be 0.70, or higher, regardless of the number of items. Under the conditions of unidimensionality, and moderately high intercorrelations, it has been recommended that four to seven items per construct (or factor) could be sufficient to reach an alpha of 0.80 [4]. For a newly developed scale, psychometric researchers have set the benchmark for alpha at 0.80 for a scale to be used [16,17].

III. THE CLASS

As the last ten years of physics education have brought reform specifically designed to improve student attitudes

about physics, the CLASS has become an important tool to assess curriculum reform. According to Google Scholar on March 20, 2014, the CLASS article [8] published in physics education research (PER) has been referenced in 253 articles. There are 62 references to it in Physical Review journals, and 34 in PER. The CLASS has been used in numerous studies published in PER (e.g., [18–20]). In addition, it has been modified for biology and chemistry [21,22] and it has been translated for use in several languages [23].

The CLASS was developed by scholars at the University of Colorado and loosely based on other established attitude and epistemological surveys towards science [8] and Fishbein's theory of attitudes [24]. Adams and colleagues also report considerable time spent interviewing experts and students in order to gain better understanding about student attitudes and beliefs about physics and learning physics. According to the creators, "This survey probes students' beliefs about physics and learning physics and distinguishes the beliefs from those of novices. The CLASS was written to make the statements as clear and concise as possible and suitable for use in a wide variety of physics courses" ([25], p. 1).

The developers of the CLASS administered it at both middle-sized multipurpose schools and large research universities in various physics courses composed of students with differing majors [8]. In addition, the researchers have reported studying the psychometric properties of the CLASS with more than 5000 students. The CLASS consists of 42 Likert-scale items probing student attitudes towards physics. Students rate their level of agreement with each item on a five-point scale from *strongly disagree* to *strongly agree*. The scoring of the CLASS is determined by the participant answers of agreement with the predetermined opinion of physics experts.

Based on factor analysis work, 26 of the 41 items were grouped into eight overlapping factors of attitudes about physics and learning physics. The eight factors determined were *Real World Connections*, *Personal Interest*, *Sense Making and Effort*, *Conceptual Connections*, *Applied Conceptual Understanding*, *Problem Solving General*, *Problem Solving Confidence*, and *Problem Solving Sophistication*. Adams *et al.* define *Real World Connections* as "physics described by the world," *Personal Interest* as "I think about physics in my life," *Conceptual Understanding* as "physics based on a conceptual framework," and *Sense Making and Effort* as "I put in the effort to make sense of physics ideas [8]." Although *Problem Solving* was not specifically defined in the paper, Sherin defines problem solving as the ability to understand the problem in relation to a particular schema and then solve the problem within that schema's techniques and equations [26]. Further understanding of the labeled categories can be found by reading the items within each category.

This eight-factor model is very complex and uses only 26 of the items administered to students. Each of the factors has at least four items and numerous items overlapping between factors, indicating constructs are not unidimensional. The factor analysis work completed on the CLASS is the first of its kind. According to the researchers, each factor was based on the following equation: $\text{Robustness} = (2cc + fl + 5|\Delta E|/\text{N}) \times 3R^2$, where cc is the average absolute value of the correlation coefficients between items, fl is the average absolute value of the factor loadings for each factor, ΔE represents the shape the scree plot, N is the number of items in the factor, and R^2 is the Pearson product moment correlation which represents how close to a straight line the scree plot is for components (eigenvalues) greater than 1 [8].

While Adams and colleagues developed an interesting approach to factor analysis that capitalizes on the strengths of both theory- and data-driven approaches, their resulting solution violates the foundation of reliability and validity: unidimensionality of each construct [11]. In a discussion about the purpose of factor analysis, Gorsuch wrote, "Each factor represents an area of generalization that is qualitatively distinct from that represented by any other factor," [27]. Yet, in the CLASS, several items are used to score more than one factor which is conceptually problematic. The categories of *Conceptual Connections, Problem Solving Confidence*, and *Problem Solving Sophistication* do not have any indicator items that are unique; all the items are scored in other categories. In addition, the researchers do not report a measure of internal reliability for the instrument or individual factors.

In a psychometric reevaluation of the chemistry version of the CLASS, Heredia and Lewis found a 16-item, three-factor solution best fit their data and Fishbein's theory of attitudes [28]. In addition, Heredia and Lewis suggest that similar psychometric reevaluations of the CLASS occur in the physics and biology versions. Their suggestion is in alignment with Crocker and Algina's argument that the "ultimate criterion for the number of factors to interpret is replicability" ([14], p. 303].

The purpose of this study is to examine the psychometric properties of the CLASS through classical test theory. In addition, we follow with Thurstone's notion of simple structure [29,30], which has been used by many researchers to support the selection of the most parsimonious factor structure that fits the data, along with Crocker and Algina's three criteria: sensibility, simple structure, and replicability [14]. The research questions are: (1) What are the psychometric properties of the CLASS on a large data set? (2) What is the empirical factor structure found from exploratory and confirmatory factor analysis?.

IV. METHODS

A. Participants and data collection

Participants include 3,844 college students enrolled in an introductory calculus-based physics course at a large

research university located in the Midwest. Typically, the course consists of mostly freshman and sophomore engineering majors, followed in number by physics and chemistry majors. Available demographic information showed the majority of students were male (77%) and either Caucasian (75.5%) or Asian American (9%). The data were randomly split into two groups, 1918 students were selected for the exploratory factor analysis and 1926 students were selected for the confirmatory factor analysis.

Data were collected pre- and postsemester over eight different academic terms starting in Spring 2006. Typical enrollment in the course is about 800 in the fall, 1200 in the spring, and about 40 students in the summer sessions. The response rate for all semesters is estimated to be slightly above fifty percent.

The survey was available online through the course web site at the beginning and end of each semester. The students volunteered to complete the online CLASS survey. There was a small extra credit incentive for students who completed both the pre- and post-survey.

B. Data analysis

Data were collected at the beginning and end of each semester. Only the predata were analyzed. This would ensure that any intervention during the course of the semester would not influence our results. The data were randomly split into two groups. Approximately half of the data were used to run a data-driven exploratory factor analysis ($n = 1918$) and the other half to confirm the newly proposed factor model ($n = 1926$). Before the statistical analyses, data cleaning was conducted on students who responded incorrectly to the monitoring item, who had the same answer for all the items, and who completed less than half of the questions. Missing data were negligible and thus handled by replacing the mean of each item for missing values.

Preliminary to conducting factor analysis, interitem correlations were conducted as recommended by Spector [10]. This provides evidence of homogeneity of construct; items that correlate relatively highly are considered to have evidence of measuring a single construct. Items that do not correlate significantly with other items that are purported to measure the same construct are problematic and should be deleted.

Exploratory factor analysis (EFA) is primarily used to determine a more parsimonious conceptual understanding of the data by identifying underlying latent constructs for the measured items which may or may not be correlated with each other [12]. EFA estimates the pattern of relationships between common factors and each measured variable in order to understand the structure of correlations among the measured variables. In other words, EFA summarizes the interrelationships between variables in order to aid in conceptualization [31].

While orthogonal rotation forces the factors to be uncorrelated, oblique rotation allows factors to be uncorrelated or correlated and is therefore a good choice in education research, as latent constructs in social science data are usually correlated to some extent [12].

The exploratory factor analysis was conducted in SPSS 18. Twenty-five of the items had elevated means (>3.5 on a 5 point scale), the items skewness ranged from −1.14 to 1.15, and kurtosis ranged from −0.93 to 1.26. Principal axis factoring was used to extract the factors along with a promax rotation for ease of interpretation, as this estimation technique is appropriate for data with skew and kurtosis less than 3 and 10, respectively. The factors were not determined to be orthogonal, thus an oblique rotation was used. In fact, cross loadings were minimized but were not entirely removed from the factor structure.

While EFA is a data-driven model, confirmatory factor analysis (CFA) is a theory-driven approach; it is used to test a hypothesis about the data [30]. Both EFA and CFA are based on a common factor model. Therefore, when examining relationships between items and constructs within a large data set, it is appropriate to randomly split the data in half, and perform an EFA with one-half, a CFA with the other. This approach provides strong support for the proposed model and is recommended as cross validation of the factor structure [12].

A 17-item factor model without cross loadings was first modeled in Amos 18 [32] to confirm the factor structure on the other half of the data. Again, since the data are categorical and not normally distributed, the asymptotically distribution-free estimation method was utilized. To improve model fit, various cross loadings and correlated variances between items of the same factor were added to the model.

V. RESULTS

The range of item response choices on the CLASS is from one to five. Initial exploratory descriptive analysis found almost half of the items had standard deviations greater than 1.0 and a few items with means close to 3.0. This indicates non-normality of the data, which is likely for the categorical data being analyzed. Table I displays the mean, standard deviation, and item-total correlations for the exploratory and confirmatory data sets. The interitem reliability for both the exploratory and confirmatory data sets with all items was high (Cronbach's α of 0.845 and 0.841, respectively). However, there were ten items (4, 7, 8, 9, 18, 19, 27, 33, 38, 41) in both data sets with all bivariate correlations with other items less than 0.275. In addition, the item-total correlation was below 0.20 for nine items in the exploratory data set and ten items in the confirmatory data set. These items were deleted because they did not correlate well with any other item; meaning they were measuring something different than other items and not representative of a single underlying construct [4].

Communality coefficients are "the variance in a measured variable the factors as a set can reproduce," ([30], p. 179). In other words, the shared variance of items. When an EFA

TABLE I. Univariate summary statistics and item-total correlations of the CLASS.

Exploratory Factor Analysis ($n = 1918$)								Confirmatory Factor Analysis ($n = 1926$)							
Item	M	SD	r_{corr}	Item	M	SD	r_{corr}	Item	M	SD	r_{corr}	Item	M	SD	r_{corr}
1	3.0	1.1	.32	22	3.0	1.0	.33	1	3.0	1.1	.31	22	3.0	1.0	.33
2	3.9	0.9	.32	23	3.9	0.8	.40	2	3.9	0.8	.31	23	3.9	0.9	.38
3	3.4	1.1	.50	24	4.0	0.8	.37	3	3.4	1.1	.49	24	4.0	0.8	.35
4	3.6	1.0	.01	25	3.3	1.0	.55	4	3.7	1.0	.05	25	3.3	1.1	.55
5	3.3	1.1	.42	26	4.1	0.7	.46	5	3.3	1.1	.43	26	4.1	0.7	.47
6	3.6	1.0	.44	27	3.2	1.1	.07	6	3.6	1.0	.41	27	3.2	1.1	.06
7	2.5	0.9	−.17	28	3.7	0.9	.43	7	2.6	0.9	−.15	28	3.7	0.9	.46
8	2.2	0.9	.15	29	4.2	0.9	.42	8	2.2	0.9	.17	29	4.2	0.8	.41
9	3.2	1.1	.15	30	4.0	0.8	.52	9	3.2	1.1	.18	30	4.0	0.7	.53
10	3.9	0.9	.37	32	3.9	0.9	.46	10	3.8	0.9	.34	32	3.9	0.9	.49
11	4.0	0.8	.44	33	3.0	1.0	−.05	11	3.9	0.9	.42	33	3.0	1.0	−.06
12	2.3	1.1	.23	34	3.8	0.8	.51	12	2.3	1.1	.25	34	3.8	0.8	.50
13	3.7	1.0	.40	35	3.9	0.9	.49	13	3.7	1.0	.37	35	3.9	0.9	.47
14	3.6	0.9	.51	36	3.1	1.0	.33	14	3.6	1.0	.49	36	3.1	1.0	.34
15	3.9	0.8	.37	37	3.4	1.1	.42	15	3.9	0.8	.37	37	3.4	1.0	.40
16	3.8	0.9	.33	38	3.6	1.0	.23	16	3.8	0.9	.30	38	3.5	1.0	.20
17	3.4	1.0	.29	39	3.8	0.8	.34	17	3.4	1.0	.27	39	3.7	0.8	.36
18	3.4	1.0	.04	40	4.0	0.8	.51	18	3.4	1.0	.02	40	4.0	0.9	.49
19	3.7	0.9	.12	41	3.2	1.0	.01	19	3.7	0.9	.09	41	3.2	1.0	−.01
20	3.9	0.9	.38	42	3.8	0.7	.50	20	3.9	0.9	.36	42	3.8	0.7	.49
21	3.4	1.1	.44					21	3.4	1.1	.44				

is performed with items of low communality, the results can be substantially distorted [11]. For the exploratory factor analysis, based on recommendations from Comrey, an iterative removal of items with low communalities (<0.30) in order to reduce distortion [33]. Typical item communalities in the social sciences are between 0.40 and 0.70, and a "high" communality is considered 0.80 or greater [19]. After each item removal, the EFA was conducted to find out whether communalities of other problematic items would increase after other poor items were removed. The remaining 30 items were analyzed and six more items were removed in an iterative manner because communalities were less than 0.30. Seven more items were removed because of they did not significantly load on any factor (>0.30), following best practices recommendations [11].

The final model was comprised of three factors consisting of 17 items. The first factor contains seven items pertaining to the student internalizing physics concepts and relating them to the world around them. This factor is called *Personal Interest and Relation to the Real World*. The second factor contains five items pertaining to student attitude towards problem solving and learning physics. This factor is called *Problem Solving and Learning*. The final factor contains five items that pertain to student level of effort in understanding physics concepts and their relationships. This factor is called *Effort and Sense Making*. The factor matrix presented in Table II shows there may be some cross loadings between factors for items 11, 21, 25, 34, and 35. Cronbach's alpha was calculated as a measure of internal reliability for each factor and for the overall assessment of students' attitudes about physics and learning physics: *Personal Interest and Relation to the Real World* $\alpha = 0.80$; *Problem Solving and Learning* $\alpha = 0.73$; *Effort and Sense Making* $\alpha = 0.69$; Overall scale $\alpha = 0.86$.

This 17-item factor model was analyzed via confirmatory factor analysis (CFA). The model fit was still not found to be within an acceptable range based on commonly accepted standards given by Hu and Bentler [34]. Items 11 and 35 were removed based on modification indices reported and the EFA results. The CFA was recalculated. Cross loadings and error covariances were added to the model to increase fit indices based on modification indices that held theoretical or conceptual justification, as recommended by Brown [35]. For example, while many cross loadings indicate poor conceptualization, we considered the modification indices suggestion to allow item 25 to cross load with the factor of *Problem Solving and Learning*. This allowed retention of the item, *I enjoy solving physics problems*, which is conceptually an important aspect of a students' attitude about physics and learning physics. In addition, it is understandable that the item would load with other items that were related to the students' own personal interest, as well as their effort in solving physics problems. This is purely for model-fitting purposes; we do not recommend, however, that this item be scored twice. We conceptualize it as an aspect of *Personal Interest and Relation to the Real World*. Items with similar wording are understood to have similar measurement error and thus were allowed to have the error variances correlated, as suggested by the modification indices.

TABLE II. Exploratory factor analysis factor structure.

	Factor 1		Factor 2		Factor 3		
	P	*S*	*P*	*S*	*P*	*S*	h^2
Factor 1: Personal Application and Relation to Real World							
3. I think about the physics I experience in everyday life.	**.716**	**.682**		.317		.272	.470
11. I am not satisfied until I understand why something works the way it does.	**.392**	**.479**	−.116	.224	.290	.422	.284
14. I study physics to learn knowledge that will be useful in my life outside of school.	**.650**	**.658**		.323		.333	.433
25. I enjoy solving physics problems.	**.607**	**.651**	.258	.464	−.161	.273	.471
28. Learning physics changes my ideas about how the world works.	**.587**	**.577**		.226		.309	.339
30. Reasoning skills used to understand physics can be helpful to me in my everyday life.	**.469**	**.582**		.327	.264	.476	.387
37. To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed.	**.587**	**.556**		.233		.238	.311
Factor 2: Problem Solving and Learning							
5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic.		.291	**.652**	**.618**		.269	.386
21. If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it.		.292	**.410**	**.505**	.169	.391	.277
22. If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations.	−.140	.142	**.514**	**.486**		.275	.251
34. I can usually figure out a way to solve physics problems.	.225	.443				.334	.371
40. If I get stuck on a physics problem, there is no chance I'll figure it out on my own.		.326	**.456**	**.581**	.253	.484	.382
Factor 3: Effort and Sense Making							
23. 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 going back through the problem.	−.121	.208	.185	.384	**.490**	**.527**	.304
24. In physics, it is important for me to make sense out of formulas before I can use them correctly.	.254	.371	−.207	.144	**.439**	**.455**	.264
29. To learn physics, I only need to memorize solutions to sample problems.		.247		.359	**.611**	**.609**	.379
32. Spending a lot of time understanding where formulas come from is a waste of time.		.325		.382	**.559**	**.608**	.375
35. The subject of physics has little relation to what I experience in the real world.	.167	.425	.189	.447	**.341**	**.522**	.333

Note: P = PatternStructure; S = FactorStructure

Results of both the 17- and 15-item model and the final fitted model are given in Table III. The chi-square value is the standard overall model goodness-of-fit (GFI) index, however, it is highly sensitive to sample size and routinely violated in CFA [36]. Therefore, structural equation modeling software provides several other ways to practically evaluate how well a hypothesized model fits the data, referred to as goodness of fit statistics [37]. There are dozens of fit statistics available for use in evaluating a model [36]. To evaluate models through the use of different evidence, researchers are recommended to report at least one fit index from each class of absolute, parsimony, and comparative fit indices [35]. In doing so, we followed Byrne's recommendations on interpretation and report the "rules of thumb" after a brief description of each fit index [37]. Root mean square residual (RMR) is an absolute fit index that is calculated based on covariance residuals, differences between observed and predicted covariances [35]. Ideally, models will have 0.05 or less RMR. The GFI is an absolute fit index where the hypothesized model is compared to the null hypothesis, in other words, the items have no structured relationship. Values close to 1.0,

TABLE III. Confirmatory factor model and fit indices.

	χ^2	df	RMR	GFI	CFI	RMSEA	BIC
3 factors with 17 items	656.84*	116	.077	.915	.628	.049	936.7
3 factors with 17 items and correlated errors	436.90*	110	.066	.944	.775	.039	762.1
3 factors with 15 items	516.70*	87	.071	.928	.675	.051	766.3
3 factors with 15 items and correlated errors	323.82*	83	.058	.955	.818	.039	603.7

*$p < .0001$. RMSEA, root-mean-square error of approximation; BIC, Bayesian information criterion.

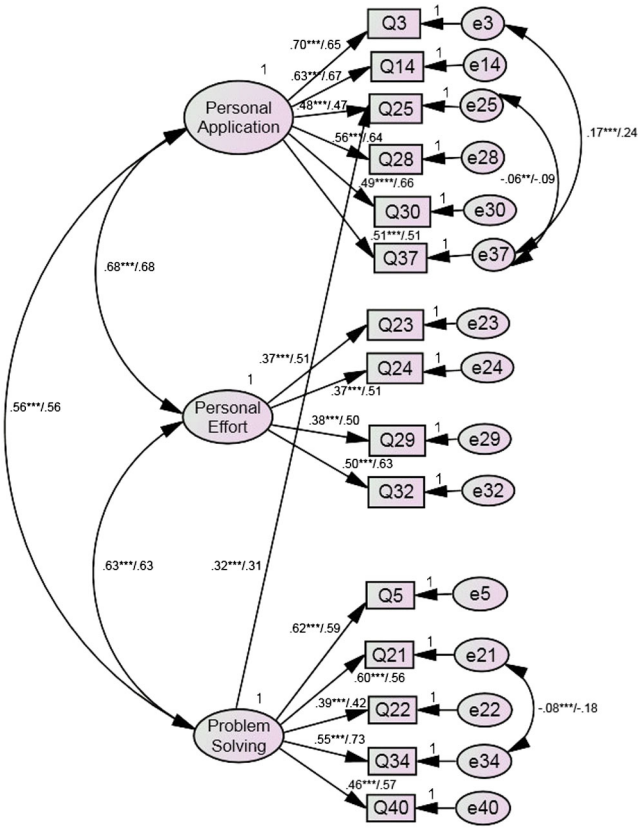

FIG. 1 (color online). Unstandardized factor loadings and covariances for confirmatory factor model. (Standardized factor loadings and correlations are listed after the slash.) "Personal Application" denotes factor 1; "Problem Solving" denotes factor 2; and "Personal Effort" denotes factor 3. **p < .01. ***p < .001.

specifically 0.95 or higher, indicate a good fit. The root-mean-square error of approximation is an index of parsimony that compares how well the hypothesized model fits the data, while accounting for the complexity of the model, and penalizes for overly complex models. This is considered one of the most widely used fit indices in CFA [37]. Values of 0.05 are indicative of a well-fitting model, whereas values of 0.08 and higher indicate misspecification. The Bayesian information criterion is a comparative index that takes into account the complexity of the model, per degrees of freedom and overall model fit. This is helpful in comparing models, as that a lower value indicates the more parsimonious model. The comparative fit index (CFI) is an incremental comparison of the hypothesized model to the base model. Values range from 0 to 1, where values of 0.95 or higher indicate excellent fits.

Consideration of the 5 indices described above led to the conclusion that the 15-item, three factor model is the best-fitting model for the data. The final model with unstandardized and standardized factor loadings is shown in Fig. 1. Factor loadings can be viewed as regression coefficients which show the strength of relationship between the items and the factors that underlie them. Internal reliability was again calculated: *Personal Application* and *Relation to the World* $\alpha = 0.82$; *Problem Solving and Learning* $\alpha = 0.73$; *Personal Effort and Sense Making* $\alpha = 0.61$; and overall scale $\alpha = 0.82$.

VI. DISCUSSION

The proposed factor structure is more parsimonious than the original factor structure and may allow for the use of a shorter measure of student attitudes towards physics and learning physics. Table IV compares the original version of the CLASS to the revised version. While several items have been removed from the original survey and there is loss of item-level data, the result is an interpretable instrument that researchers can use to understand student attitudes. All of the items in the proposed version are scored, so all of the questions students answer have utility. Additionally, the proposed factor structure has strong psychometric properties which support that each factor is indeed measuring one aspect of student attitudes about physics. Although the CFI did not meet the level of a superior fitting model, taken into account with the other fit indices, the final model is the best fit for these data. From a developmental perspective, the next step is to empirically test the newly created shortened

TABLE IV. The Original and Proposed Revision to the CLASS Categories and Survey Items.

Original CLASS [14]		Proposed Revision	
Categories	Survey Items	Categories	Survey Items
Real World Connection	**28**, **30**, 35, 37	Personal Application and Relation to Real World	3, 14, 25, 28, 30, 37
Personal Interest	**3**, **11**, 14, **25**, 28, **30**	Problem Solving/Learning	5, 21, 22, 34, 40
Sense Making/Effort	**11**, 23, 24, **32**, 36, 39, **42**	Effort/Sense Making	23, 24, 29, 32
Conceptual Connections	**1**, **5**, **6**, **13**, **21**, **32**		
Applied Conceptual Understanding	**1**, **5**, **6**, 8, **21**, **22**, **40**		
Problem Solving General	**13**, **15**, **16**, **25**, 26, **34**, **40**, **42**		
Problem Solving Confidence	**15**, **16**, **34**, **40**		
Problem Solving Sophistication	**5**, **21**, **22**, **25**, **34**, **40**		
Not Scored	4, 7, 9, 31, 33, 41		

Note.—Items in bold are scored more than once.

version. Future research should consider whether additional revisions to the instrument may improve the theoretical understanding of the items and improve model fit.

When the models of the exploratory and confirmatory factor analysis were compared with the original factor structure proposed by Adams *et al.* [8], numerous similarities were found between the determined factors. The first factor, *Personal Application* and *Relation to the World*, is the combined factors of *Personal Interest* and *Real World Connection* without items 11 and 35, which did not have strong factor loadings in the confirmatory model but were included in the exploratory factor model. Further analysis with the postdata and other data sets would help explain this incongruence with the two items.

The second factor, *Problem Solving and Learning*, has the exact same items as the original *Problem Solving Sophistication* factor. This suggests a more parsimonious factor of problem solving and may indicate that the previous three factors of problem solving are better described by the single underlying construct of problem solving. The third factor, *Personal Effort and Sense Making*, is contained in the original *Sense Making and Effort* factor. However, there are four items in the original factor not retained in the exploratory analysis completed here. Item 29 ("To learn physics, I only need to memorize solutions to sample problems") was not included in the original factor structure. Item 29 seems an appropriate addition to a factor including other items referencing sense making and effort. Theoretically, experts understand physics on a deeper level than simple memorization of solutions, whereas novices may approach learning physics by memorization.

In addition to the similarities this factor structure shares with the original CLASS factor structure, our results find similar factors as those found by Heredia and Lewis in the chemistry version of the CLASS [28]. Their research also found factors of Personal Interest and Problem Solving. In the chemistry version of CLASS, there are also unique items related to Atomic-Molecular Perspective of Chemistry. This provides further evidence that the CLASS is coming closer to a stable factor structure that can be used in different settings. One caveat to this finding, however, is that the student demographics in this study were largely white males and it is unknown whether this factor structure would hold with students from other demographics. Heredia and Lewis had a somewhat more diverse student sample, with white males comprising 42% and whites 57%. In keeping with Messick's proposition that assessments should be viewed from a developmental perspective, as researchers consider student attitudes in populations with larger variation in student demographics, this factor structure should again be examined specifically with underrepresented minority groups [5]. In addition, use of item response theory could further examine whether students from diverse groups understand the CLASS in the same way.

The finding that the three factors of *Personal Application and Relation to the World, Problem Solving and Learning*, and *Personal Effort and Sense Making*, correlate significantly with each other and have a high level of overall internal reliability, suggests that a general attitude about physics is emerging from the data. The proposed factor structure will enable researchers to parse out the different attitudes from each other, but still appreciate that student overall attitude may be attributing to how interested they personally are in physics, the way they are able to relate physics to their world, and how much effort they are willing to devote to learning. The internal reliability of factors *Problem Solving and Learning*, and *Personal Effort and Sense Making* are still below what is ideal and future revisions of CLASS should consider what other aspects of those constructs in relation to Fishbein's theory [24] could be added to more accurately measure students' attitudes around problem solving and exerting effort in physics.

VII. FUTURE DIRECTIONS

From a use perspective, the proposed version of the CLASS will enable researchers to focus on assessing specific areas rather than trying to make sense out of interpreting overlapping constructs. In addition to providing more targeted research on student attitudes related to physics, a revision of the CLASS has potential to be of more practical use for instructors to evaluate the areas that need to be addressed more purposefully in courses. A 15-item survey is relatively quick for an instructor to administer and could be done midsemester to assess how well students are maintaining or improving in how they relate physics to the world, problem solve, and put forth effort in learning. Instructors could then adjust lectures to more clearly address the areas in which students are struggling. The findings from the end of the course assessment along with the revised CLASS could be used to improve future offerings of the course.

In order for the science education community to capitalize on the "opportunity" that Pellegrino [1] calls for in increasing the rigor of assessments, researchers must not only employ rigorous methods for developing and testing assessments, but researchers must continuously re-examine the psychometric properties of popular assessments. Validity is an ongoing process that cannot be separated from its context [5]. Therefore, in order to have confidence in an assessment's results, the original psychometric properties should be replicated in additional studies on different groups, and, if not found, modifications to the assessment, based on empirical findings that are theoretically coherent, should be shared with the community. Through this developmental perspective of assessments, the science education research community will have greater confidence in the validity of their results and the interventions designed to increase students' outcomes.

Researchers have a responsibility to be informed users when they are considering the best way to determine the outcomes of curricular reform. While psychometric studies may not be, to some, as interesting as the results of innovative curriculum reform, what will happen in curricular reform will, at least in part, be influenced by assessment. Care should be taken in how we conceptualize and measure students' attitudes and the evidence we have to support our measures. Through collaboration with those who have measurement and assessment expertise, physics education researchers can become more informed of critical assessment issues and develop more robust measures.

[1] J. W. Pellegrino, Assessment of science learning: Living in interesting times, J. Res. Sci. Teach. **49**, 831 (2012).

[2] American Educational Research Association, American Psychological Association, National Council on Measurement, & Joint Committee on Standards for Educational and Psychological Testing (U.S.) (1999), *Standards for Educational and Psychological Testing* (American Educational Research Association, Washington, DC, 1999).

[3] T. Kubiszyn and G. Borich, *Educational Testing and Measurement*, 8th ed. (Wiley, New York, 2007).

[4] R. G. Netemeyer, W. O. Bearden, and S. Sharma, *Scaling Procedures: Issues and Applications* (SAGE, Thousand Oaks, CA, 2003).

[5] S. Messick, Validity of psychological assessment, Am. Psychol. **50**, 741 (1995).

[6] T. Hong, S. Purzer, and M. E. Cardella, A psychometric re-evaluation of the design, engineering and technology survey, J. Eng. Educ. **100**, 800 (2011).

[7] J. A. Arjoon, X. Xu, and J. E. Lewis, Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, J. Chem. Educ. **90**, 536 (2013).

[8] 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. ST Phys. Educ. Res. **2**, 010101 (2006).

[9] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research* 2nd ed. (Sage, Thousand Oaks, CA, 2007).

[10] P. E. Spector, *Summated Rating Scale Construction: An Introduction* (Sage, Thousand Oaks, CA, 1991).

[11] J. C. Nunnally and I. H. Bernstein, *Psychometric Theory*, 3rd Edition. (McGraw-Hill, New York, 1994).

[12] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, Evaluating the use of exploratory factor analysis in psychological research, Psychological Methods **4**, 272 (1999).

[13] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[14] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Wadsworth Publishing Co., Cengage Learning, Stamford, CT, 2006).

[15] J. M. Cortina, What is coefficient alpha? An examination of theory and applications, J. Appl. Psych. **78**, 98 (1993).

[16] L. A. Clark and D. Watson, Constructing validity: Basic issues in objective scale development, Psychol. Assess. **7**, 309 (1995).

[17] J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, Criteria for scale selection and evaluation, *Measures of Personality and Social Psychological Attitudes* **1**, 1 (1991).

[18] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

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

[20] E. Brewe, A. Traxler, J. de la Garza, and L. H. Kramer, Extending positive CLASS results across multiple instructors and multiple classes of modeling instruction, Phys. Rev. ST Phys. Educ. Res. **9**, 020116 (2013).

[21] K. Semsar, J. K. Knight, G. Birol, and M. K. Smith, The Colorado Learning Attitudes about Science Survey (CLASS) for use in biology, CBE Life Sci. Educ. **10**, 268 (2011).

[22] 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).

[23] CLASS, www.colorado.edu/sei/class (n.d.).

[24] I. Ajzen and M. Fishbein, Attitude-behavior relations: A theoretical analysis and review of empirical research, Psychol. Bull. **84**, 888 (1977).

[25] W. K. Adams, K. K. Perkins, M. Dubson, N. D. Finkelstein, and C. E. Wieman, The design and validation of the Colorado Learning Attitudes about Science Survey, *PERC* (2004).

[26] B. L. Sherin, How students understand physics equations, Cognit. Instr. **19**, 479 (2001).

[27] R. L. Gorsuch, *Factor Analysis*, 2nd ed. (Lawrence Erlbaum, Mahwah, New Jersey, 1983).

[28] 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).

[29] L. L. Thurstone, *The Vectors of Mind* (University of Chicago Press, Chicago, 1935).

[30] B. Thompson, *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, 1st ed. (Am. Psych.l Assoc., Washington, DC, 2004).

[31] R. L. Gorsuch, Common factor analysis versus component analysis: Some well and little known facts, Multivariate Behav. Res. **25**, 33 (1990).

[32] IBM., *White Paper: Structural Equation Modeling with IBM SPSS Amos* (2005).

[33] A. L. Comrey, Common methodological problems in factor analytic studies, J. Consult. Clin. Psychol. **46**, 648 (1978).

[34] L. Hu and P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Equation Modeling: A Multidisciplinary Journal **6**, 1 (1999).

[35] T. A. Brown, *Confirmatory Factor Analysis for Applied Research* (Guilford Press, New York, 2012).

[36] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 3rd ed. (Guilford Press, New York, 2011).

[37] B. M. Byrne, *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming* (Routledge, New York, 2013).