Exploring the CLASS with Item Response Theory

Elaine Christman, Paul Miller, and John Stewart
Department of Physics and Astronomy, West Virginia University,
135 Willey St., Morgantown, West Virginia, 26506, United States

This work applied exploratory factor analysis (EFA) and graded Item Response Theory (IRT) to a large sample ($N = 4522$) of Colorado Learning Attitudes about Science Survey (CLASS) post-test scores. EFA failed to reproduce the factor structure suggested by the authors of the CLASS and strongly supported the alternate 3-factor model suggested by Douglas et al. Graded IRT allowed an examination of the progression from non-expert-like to expert-like beliefs. This progression was generally uniform with a linear relation between the difficulty of each step in the progression. Some items within the factors identified by Douglas et al. had difficulty and discrimination parameters substantially different from other items in the factor suggesting the subscale is not unidimensional. The expert-like latent ability trait estimated by IRT correlated more strongly with measures of physics performance than measures of general academic performance indicating that expert-like beliefs are not a general property of high performing students.

2020 PERC Proceedings edited by Wolf, Bennett, and Frank; Peer-reviewed, doi.org/10.1119/perc.2020.pr.Christman

I. INTRODUCTION

Since its introduction in 2006, the Colorado Learning Attitudes about Science Survey (CLASS) has become the most widely applied instrument in Physics Education Research (PER) to measure a student's expert-like beliefs about physics [1]. Items on the CLASS are scored on a three-level scale (non-expert-like, neutral, and expert-like) determined by a comparison with the responses of an expert panel of physicists. The percentage of expert-like responses typically decreases during a first-semester physics course [2]. Chinese secondary students also showed an overall decline in expert-like beliefs over a longer time scale [3]. Gray *et al.* [4] asked students to complete a modified CLASS that solicited students' personal attitudes and their beliefs about how physicists would respond to the same questions. Students' ideas about physicists' beliefs were quite stable over the course of a semester, but their personal beliefs were most often negatively affected by instruction. Specific pedagogical approaches, including Physics by Inquiry [5]; Peer Instruction [6]; Physics and Everyday Thinking [7]; and Modeling Instruction [8], have been shown to result in attitudinal shifts toward expert-like thinking.

The CLASS has been used in a broad variety of studies. Kost *et al.* used CLASS scores to explore gender differences in conceptual post-test scores [9]. Ding found a significant causal relationship between pretest scores on the CLASS and normalized gains on a mechanics conceptual inventory [10]. Traxler and Brewe disaggregated CLASS data to examine the effects of Modeling Instruction on the attitudes of women and underrepresented minorities [11]. Baily and Finkelstein used the CLASS to monitor students' development of a probabilistic quantum-mechanical perspective in response to modern physics instruction [12]. Gire *et al.* used the CLASS to compare the views of introductory physics students majoring in physics and engineering [13]. The physics CLASS has been used as a basis for the development of similar surveys for chemistry [14], biology [15], laboratory practices [16], and informal science education experiences [17].

The CLASS consists of 42 statements with which a student may strongly agree, agree, remain neutral, disagree, or strongly disagree. The authors of the CLASS reported that exploratory factor analysis suggested eight factors, each consisting of four to eight items [1]. Twenty-seven of the items loaded onto a factor, with 16 of these 27 items loading onto two or more factors. Nine items did not load onto a factor and six additional items were not scored. For each of the scored items, a panel of experts rated whether high scores or low scores on the item represented expert-like beliefs, producing a 3-level scale: non-expert-like, neutral, and expert-like. The overall instrument and each subscale (factor) is then given a score representing the percentage of expert-like response (% favorable) [1]. Douglas *et al.* were unsuccessful in replicating this factor structure [18] and reported that a three-factor model offered acceptable fit statistics. A modified instrument was proposed containing only the 15 items contained in their subscales [18]. Sawtelle *et al.* found that students at a predominately Hispanic university overwhelmingly interpreted the statements as intended and flagged one item, item 21, as misinterpreted by over one third of students [19].

The CLASS is widely used in PER to evaluate the impact of reformed curricula so it is valuable to probe what the CLASS truly measures and whether these attitudes toward physics predict physics achievement. This work seeks to answer 3 research questions: *RQ1: What factor structure is identified for the CLASS? RQ2: How can Item Response Theory (IRT) be used to further understand the CLASS? RQ3: What academic performance measures most strongly correlate with the latent expert-like ability trait estimated by IRT?*

II. METHODS

The data for this study were collected at an eastern land-grant university serving approximately 30,000 students. The undergraduate demographics of the university were 80% white, 6% international, 4% African-American, 4% Hispanic, 2% Asian, 4% two or more races, and other groups less than 1%. The students were enrolled in an introductory calculus-based mechanics course serving primarily engineering and physical sciences majors. A total of 4522 CLASS post-test records form the sample for this study.

Item Response Theory (IRT) is composed of a broad collection of statistical models for the response patterns to different types of examinations. In general, IRT models predict the probability a student selects the correct answer or a given response on a multiple-choice instrument in terms of a latent trait θ which measures a student's general ability to correctly answer items in the instrument. The CLASS, when graded as suggested by the authors, classifies a student's response to an item with three levels: non-expert-like, neutral, and expert-like. These levels are coded as -1, 0, and 1 in the dataset, respectively. The probability a student is classified as on one of these levels on can be modeled by the "graded" model from IRT which estimates two probability functions: $P(x \geq 0|\theta)$ and $P(x = 1|\theta)$ where θ is a latent trait modeling a student's expert-like ability. Because the probability must sum to one for each item, the probability of selecting each level of expert-like-ness can be calculated from these two probabilities: $P(x = -1|\theta) = 1 - P(x \geq 0|\theta)$, $P(x = 0|\theta) = P(x \geq 0|\theta) - P(x = 1|\theta)$.

Each probability is modeled using the 2-parameter logistic (2PL) function. An individual difficulty d_k is estimated for each probability function where $k = 1$ or 2 indexing the 2PL models (Eqns. 1 and 2). The difficulty (similar to as it is defined in Classical Test Theory [20]) is larger for items for which there is a higher proportion of expert-like responses, counter to what would be intuitively expected. A single discrimination a is estimated for both probability functions; the discrimination measures how well the item differentiates high and low expert-like ability (θ) students. The probability functions are shown in Eqns. 1 and 2 and represent the probability

P_{ij} that student i with ability θ_i selects response x of item j.

$$P_{ij}(x \geq 0|\theta) = \frac{\exp[a_j \cdot \theta_i + d_{1j}]}{1 + \exp[a_j \cdot \theta_i + d_{1j}]}, \quad (1)$$

$$P_{ij}(x = 1|\theta) = \frac{\exp[a_j \cdot \theta_i + d_{2j}]}{1 + \exp[a_j \cdot \theta_i + d_{2j}]}, \quad (2)$$

where a_j is the item discrimination and d_{1j} and d_{2j} the item difficulties.

The exponential nature of Eqn. 1 and 2 can make the 2PL functions challenging to interpret intuitively. Qualitatively they produce S-shaped curves which approach 0 as $\theta \to -\infty$ and 1 are $\theta \to +\infty$; students of very low expert-like ability θ have a very low probability of answering the neutral or expert-like responses, conversely students with high θ have a high probability of answering the expert-like option. At $\theta_k = -d_k/a$, $P(\theta_k) = \frac{1}{2}$ where $k = 1$ or 2. The ability where the probability is one-half seems a natural point to select as the transition from the less expert-like state to the more expert-like state. Note, θ approximately follows a standard normal distribution with mean 0 and standard deviation 1.

III. RESULTS

A. Factor Analysis

Exploratory factor analysis was performed on the CLASS and showed that three factors were optimal for this sample using all items scored in Adams *et al.* [1]. These three factors were very similar to those identified by Douglas *et al.* [18]. We further confirmed the factor analysis structure of Douglas *et al.* by performing factor analysis on only the fifteen items retained at the end of their 2014 paper. Our data reproduced the same three factors in both our pretest and post-test datasets. These factors are called the "Douglas subscales" in this work. Douglas named these factors *Personal Application (and Relation to the Real World)* (PARRW), consisting of items 3, 14, 25, 28, 30, and 37; *Problem Solving and Learning* (PSL) consisting of items 5, 21, 22, 34, 40; and *Effort and Sense Making* (ESM) consisting of items 23, 24, 29, 32.

The internal reliability of these subscales was characterized with Cronbach's α. For the post-test, PARRW had $\alpha = 0.76 \pm 0.01$, PSL $\alpha = 0.66 \pm 0.02$, and ESM $\alpha = 0.60 \pm 0.02$. The α for the second and third subscale both fall below the usual threshold of 0.70 for low-stakes testing. The 36-item instrument (omitting only the items not scored in Adams *et al.*) had $\alpha = 0.86 \pm 0.01$, near the threshold for high stakes testing, showing the instrument may be better characterized as measuring a single construct.

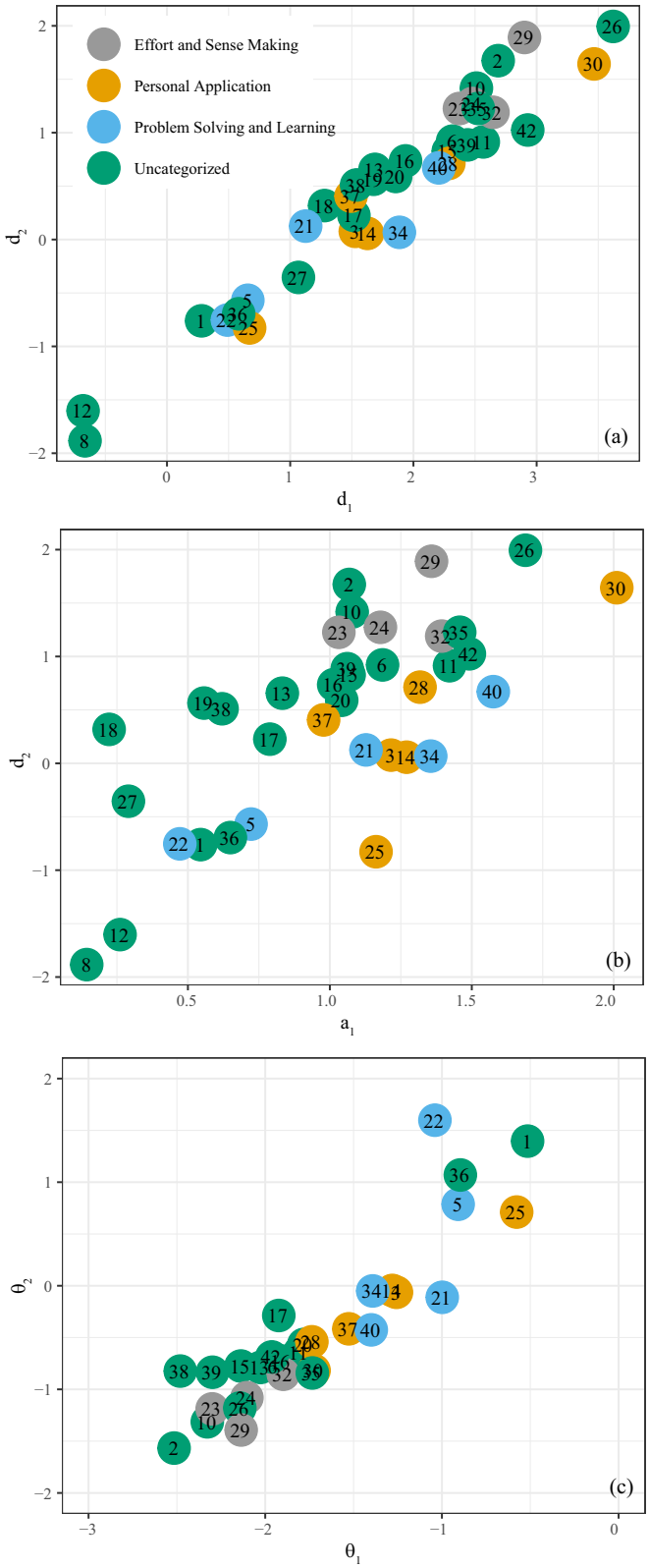

FIG. 1. Figure (a) plots d_2 vs. d_1, Fig. (b) d_2 vs. a, and Fig. (c) θ_2 vs. θ_1.

B. IRT Results

The results of fitting the graded IRT model are shown in Fig. 1. Figure 1(a) plots the two difficulties d_2 versus d_1, showing a fairly linear relationship. Fitting a linear model to this data yielded a slope of 0.88 and intercept of -1.1 with $R^2 = 0.93$. As expected, d_1 was larger than d_2; it is less probable that the student selects the expert-like response than the neutral or non-expert-like response. The strong linear relation shows the ability needed to progress from non-expert-like to neutral to expert-like is fairly consistent for all items. Figure 1(b) plots d_2 against a. Because of the linear relation of d_2 and d_1, a plot of d_1 and a is similar but shifted by the difference in d_1 and d_2,

Figure 1(c) plots θ_2 vs. θ_1. For reference, θ_1 is the ability at which the probability of selecting the non-expert-like response equals the probability of selecting the neutral or expert-like response and θ_2 is the ability at which the probability of selecting either the non-expert-like or neutral response equals the probability of selecting the expert-like response. Five items are not shown in Fig. 1(c) because they had either anomalously low or high values of θ_1 or θ_2; all were uncategorized. Items 18, 19, and 27 had very low θ_1 indicating almost no student selected the non-expert-like response (item 18, $\theta_1 = -5.7$, $\theta_2 = -1.4$; item 19, $\theta_1 = -3.7$, $\theta_2 = 1.2$; item 27, $\theta_1 = -3.0$, $\theta_2 = -1.0$). Item 18 ("There could be two different correct values to a physics problem if I use different approaches") and item 27 ("It is important for the government to approve new scientific ideas before they can be widely accepted") have very small θ_1 and small θ_2, virtually no student selects the non-expert-like response and most students select the expert-like response; these items provide little information about the student's expert-like thinking and probably should be removed from the instrument. Item 19 ("To understand physics I discuss it with friends and other students") has very small θ_1 but a relatively large θ_2; most students select the neutral response while very few students select the expert-like response. This item seems a matter of personal learning mode that could be influenced by the use of interactive engagement methods in both lab and lecture. Items 8 and 12 have anomalously large θ_k (item 8, $\theta_1 = 2.6$, $\theta_2 = 6.1$; item 19, $\theta_1 = 4.6$, $\theta_2 = 13$); almost no students select a response other than the non-expert response for these items. Item 8 ("When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values") while not optimal, certainly captures much of many students' experience with problem solving in physics; that most students agree with this statement seems natural. Item 12 ("I cannot learn physics if the teacher does not explain things well in class") equally seems like a statement to which a student in a physics class would naturally respond positively. Both items 8 and 12 contribute virtually no additional descriptive power to the instrument and should be removed. Gray *et al.* found that students were least successful at identifying "physicist" responses for these two items, and that item 12 had the lowest consistency among experts [4]. Reyes and Rakkapao removed item 8 from their Item Response Curve analysis; 75% of students in their sample chose the non-expert-like response, and an expert-like response on this item correlated with less expert-like attitudes overall [21].

Beyond the problematic items discussed above, two additional uncategorized items stand out in Fig. 1(c): items 1 and 36. Both have relatively higher θ_1 than other items (the non-expert-like response is selected somewhat more often than other items) and high θ_2 (few students select the expert-like response). These items represent behaviors that often differentiate more expert-like students from students applying more novice methods: an over-reliance on memorization (item 1, "A significant problem in learning physics is being able to memorize all the information I need to know") and the habit of checking understanding by working a problem with an alternate method (item 36, "There are times I solve a physics problem more than one way to help my understanding"). Neither item appeared in a subscale in the original CLASS study [1]. These items seem to represent the beginning of an "Expert-like Habits" subscale missing from the instrument which could include items on checking units, drawing a diagram, sketching the result of a calculation, and including textual reasoning.

TABLE I. Average subscale discrimination, difficulty, θ_k, and score.

Subscale	a	d_1	d_2	θ_1	θ_2	Score
PARRW	1.33	1.85	0.34	-1.39	-0.26	0.34
PSL	1.05	1.27	-0.09	-1.21	0.08	0.19
ESM	1.24	2.60	1.40	-2.10	-1.12	0.64

Items from the three subscales identified in Douglas *et al.* [18] largely cluster together in Fig. 1(c) with some notable exceptions. Table I presents the average of the discrimination, difficulty, θ_k, and item scores for each subscale. The Effort and Sense Making subscale has consistently higher d_k leading to lower θ_k than other subscales; as such, these items cluster in the lower left quadrant of Fig. 1(c). These items represent novice-like attitudes toward problem solving; it is comforting that a strong majority of students have at least a neutral attitude with the majority of students having an expert-like attitude. These items also have strong discrimination (Fig. 1(b)) between expert-like and non-expert-like students.

Most items in the Personal Application and Problem Solving and Learning subscales cluster around the point $\theta_1 = -1.5$ and $\theta_2 = -0.25$ in Fig. 1(c) indicating that the majority of the students have at least a neutral attitude toward these constructs with students fairly equally split between the neutral and expert-like responses. Three items are well separated from these clusters in Fig. 1(c): items 5, 22, and 25. Items 5 and 22 of the Problem Solving and Learning subscale involve the transfer of problem solving skills or the initial application of newly learned skills. The other items in the scale items 21, 34, and 40 involve general problem solv-

TABLE II. Correlation Table

	θ	%Expert-like Pre	%Expert-like Post	PARRW Post Score	ESM Post Score	PSL Post Score
Expert-like ability (θ)		0.51	0.96	0.78	0.65	0.70
FMCE Post %	0.36	0.29	0.37	0.23	0.11	0.39
Physics Test Average	0.37	0.29	0.39	0.20	0.14	0.47
Physics Grade	0.28	0.24	0.30	0.13	0.10	0.35
Cumulative College GPA	0.14	0.12	0.17	0.03	0.10	0.14
High School GPA	0.09	0.14	0.11	-0.03	0.13	0.12
ACT/SAT Math Percentile	0.14	0.20	0.16	0.03	0.12	0.23
ACT/SAT Verbal Percentile	0.17	0.25	0.20	0.02	0.14	0.21

ing self-efficacy. The strong differences in θ_k and discrimination suggest this scale may not be unidimensional and that additional item development could separate the items into two separate subscales. Likewise, item 25 of the Personal Application subscale, which simply states "I enjoy solving physics problems," has little obvious connection to other items in the scale or to other items in the instrument. It seems likely that item 25 factors with the subscale because similar items are not present in the instrument and that additional development could produce an "Enjoyment of Physics" subscale.

C. Correlates of the Latent Trait

Correlations among the latent trait θ, students' pretest and post-test percentages of expert-like responses, scores on the three Douglas subscales, FMCE post-test scores, physics test averages, college GPA, and high school achievement measures are shown in Table II. The latent trait θ estimated by the CLASS correlated much more strongly with measures of physics success (FMCE scores, physics test average, and physics grade) than with other measures of academic achievement (GPA and ACT/SAT percentiles), suggesting the CLASS expert-like physics attitudes are related to physics achievement, not simply a characteristic of more academically prepared students. Of the three subscales, Personal Application correlates most strongly with θ, while the Problem Solving and Learning subscale correlates most strongly with FMCE score, physics test average, and physics grade.

IV. DISCUSSION AND CONCLUSIONS

This study sought to address three research questions. *RQ1:* This study supported the identification by Douglas *et al.* of a three-factor solution to the CLASS and reproduced their suggested subscales [18].

RQ2: The graded IRT model allowed the identification of an overall discrimination and two levels of difficulty for each item and allowed the estimation of a latent expert-like ability, θ, for each student. The difficulty and discrimination provided a fine-grained picture of the transition from non-expert-like to neutral to expert-like beliefs. Multiple problematic items not part of the Douglas subscales (uncategorized items) were identified (items 8, 12, 18, 19, and 27); these items were answered as expert-like or non-expert-like by the majority of the students. Figure 1 shows that items within the Effort and Sense Making subscale generally cluster together in all plots. The other two subscales have clear outlying items, particularly in Fig. 1(c). This suggests that neither the Personal Application nor the Problem Solving and Learning subscales are unidimensional. We suggest that these the outlying items could be developed into independent subscales.

RQ3: The latent expert-like ability θ was more strongly correlated with measures of physics achievement than with general measures of high school or college achievement. The Problem Solving and Learning subscale was more strongly correlated with physics achievement, a medium effect, than either the Effort and Sense Making or the Personal Application subscales, both small effects. This latent expert-like ability trait provides a student-level measure of expert-like belief which should allow a more nuanced understanding of the instrument in future studies.

The factor analysis results strongly support the Douglas subscales and suggest that the eight factors proposed by the authors of the CLASS are not a useful tool for interpreting student results. The results of graded IRT analysis could be used to begin development of the next generation of physics attitudinal survey, discarding problematic items that do not provide much meaningful information about students' transition to expert-like thinking and developing new items for the proposed physics enjoyment and expert-like habits subscales. This research reveals that expert-like attitudes on the CLASS do indeed correspond to measures of success in introductory mechanics. Further work will also explore students' pretest to post-test transitions in expert-like beliefs.

This work was supported by the National Science Foundation under Grant Nos. EPS-1003907 and ECR-1561517.

[1] W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. Wieman, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).
[2] A. Madsen, S. McKagan, and E. Sayre, Phys. Rev. ST Phys. Educ. Res. **11**, 010115 (2015).
[3] P. Zhang and L. Ding, Phys. Rev. ST Phys. Educ. Res. **9**, 010110 (2013).
[4] K. Gray, W. Adams, C. Wieman, and K. Perkins, Phys. Rev. ST Phys. Educ. Res. **4**, 020106 (2008).
[5] B. A. Lindsey, L. Hsu, H. Sadaghiani, J. W. Taylor, and K. Cummings, Phys. Rev. ST Phys. Educ. Res. **8**, 010102 (2012).
[6] P. Zhang, L. Ding, and E. Mazur, Phys. Rev. Phys. Educ. Res. **13**, 010104 (2017).
[7] V. K. Otero and K. E. Gray, Phys. Rev. ST Phys. Educ. Res. **4**, 020104 (2008).
[8] E. Brewe, A. Traxler, J. de la Garza, and L. H. Kramer, Phys. Rev. ST Phys. Educ. Res. **9**, 020116 (2013).
[9] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Physical Review Special Topics-Physics Education Research **5**, 010101 (2009).
[10] L. Ding, Phys. Rev. ST Phys. Educ. Res. **10**, 023101 (2014).
[11] A. Traxler and E. Brewe, Phys. Rev. ST Phys. Educ. Res. **11**, 020132 (2015).
[12] C. Baily and N. Finkelstein, Phys. Rev. ST Phys. Educ. Res. **5**, 010106 (2009).
[13] E. Gire, B. Jones, and E. Price, Phys. Rev. ST Phys. Educ. Res. **5**, 010103 (2009).
[14] W. K. Adams, C. E. Wieman, K. K. Perkins, and J. Barbera, Journal of Chemical Education **85**, 1435 (2008).
[15] K. Semsar, J. Knight, G. Birol, and M. Smith, CBE-Life Sciences Education **10**, 268 (2011).
[16] B. Zwickl, N. Finkelstein, and H. Lewandowski, AIP Conference Proceedings **1513**, 442 (2013).
[17] R. Wulf, L. Mayhew, and N. Finkelstein, AIP Conference Proceedings **1289**, 337 (2010).
[18] K. A. Douglas, M. S. Yale, D. E. Bennett, M. P. Haugan, and L. A. Bryan, Phys. Rev. ST Phys. Educ. Res. **10**, 020128 (2014).
[19] V. Sawtelle, E. Brewe, and L. Kramer, Phys. Rev. ST Phys. Educ. Res. **5**, 023101 (2009).
[20] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).
[21] M. Reyes and S. Rakkapao, Eur. J. Phys. **41**, 045703 (2020).