Analyzing the dimensionality of the Energy and Momentum Conceptual Survey using Item Response Theory

Xian Wu and Matthew W. Guthrie
Department of Physics, University of Connecticut, Storrs, CT, 06269

Erin M. Scanlon
Department of Physics, University of Connecticut - Avery Point, Groton, CT, 06340

Yaoguang Li
AGQ Solutions, South Windsor, CT, 06074

There have been myriad conceptual inventories developed and used by the physics education research community (e.g., FCI, FMCE, BEMA, CSEM). However, continued examination of the psychometric properties of these inventories is important for improving inventory development and usage practices. In this study, we investigate the dimensionality of the Energy and Momentum Conceptual Survey (EMCS) test items and explore the structure of students' understanding of these concepts (i.e., whether the student response data showed multiple distinct test traits). To investigate the psychometric properties of this survey, we surveyed a sample of 253 students participated in the pretest and 201 students for the posttest assessments. Statistical analyses, guided by Item Response Theory (IRT), were conducted using R software. The results of Bootstrap modified parallel analysis tests (BMPAT) revealed significant differences between unidimensional models and the actual data, indicating the presence of multidimensionality (i.e., dimensions correspond to the physics concepts/abilities being tested) in the EMCS test items. Exploratory analyses suggested that a 2 or 3-dimensional model best fits the data. However, categorizing items based on designated concepts did not improve the model fit. The findings imply that students may interpret the items differently than intended by the test designers.

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

I. INTRODUCTION AND BACKGROUND

Since its introduction nearly 40 years ago, the Force Concept Inventory (FCI) [1] has spurred numerous studies aimed at developing and analyzing research-based conceptual inventories which are generally used to measure student learning in physics courses [2–4]. In addition to many assessments focusing on topics other than forces, several alternative assessments to the FCI have been developed, including the Energy and Momentum Conceptual Survey (EMCS) [5], The EMCS is a 25-item multiple-choice inventory specifically designed to evaluate students' understanding of energy and momentum concepts in algebra-based and calculus-based introductory physics courses. This study examines data collected during the Fall 2022 and Spring 2023 semesters from a calculus-based Introductory Physics course designed for Introductory Physics for Life Science students (IPLS) and taught using the Studio physics approach. Our aim was to investigate students' interpretations on the testing items and their understandings of energy and momentum concepts based on their responses. The analyses in this study were guided by Rasch Measurement Theory [6] for the alignment of the data (i.e., student responses) and the proposed statistical models (i.e., from one-parameter Item Response Theory). In the sections below, we will introduce the survey and the psychometric theories underpinning our analysis.

A. The Energy and Momentum Conceptual Survey

The 25 multiple-choice items in the EMCS were developed based on a study conducted at the University of Pittsburgh, which utilized free-response questions and think-aloud protocols [5]. The authors of the survey reported reasonable reliability coefficients [7] based on the results obtained from graduate and undergraduate students. Notably, numerous studies have acknowledged the soundness of the EMCS construction process [8–10] and the overall quality of the EMCS. However, to the authors' knowledge, limited research has been conducted to investigate student performance on the EMCS.

B. Instructional Context

At the participating institution, the physics department implemented a Studio physics program for all calculus-based introductory physics courses on the main campus starting from the Fall 2019 semester. The Studio physics program primarily follows the SCALE-UP model [11], which promotes interactive and collaborative learning. For the course sequence offered to students in life science and pre-medical programs, the first author adopted an Introductory Physics for Life Sciences (IPLS) perspective, utilizing relevant resources available on the Living Physics Portal [12] for course design. The computer science program at the participating institution accepts a course sequence that is equivalent to the traditional engineering physics sequence. Consequently, the course sequence has generated consistent interest among computer science students.

II. METHODS

A. Item Response Theory Models

Item Response Theory (IRT) models a person's response to a question as a result of both the person's characteristics and the question's properties [13]. In the unidimensional IRT framework, each student ν is associated with a proficiency parameter θ_ν, and each item i is associated with a difficulty parameter b_i. The probability P that student ν responds correctly to item i (i.e., $x_{\nu_i} = 1$) is expressed using a logistic function:

$$P(x_{\nu_i} = 1) = \frac{\exp(\theta_\nu - b_i)}{1 + \exp(\theta_\nu - b_i)}. \quad (1)$$

Equation (1) represents the one-parameter logistic (1PL), also known as the Rasch model. In addition to the difficulty parameter b_i, the two-parameter (2PL) IRT model includes a discrimination parameter a_i:

$$P(x_{\nu_i} = 1) = \frac{\exp(a_i(\theta_\nu - b_i))}{1 + \exp(a_i(\theta_\nu - b_i))}. \quad (2)$$

a_i is a positive parameter that indicates an item's ability to differentiate among students, although extremely high values are uncommon. The three-parameter (3PL) model further includes a pseudo-guessing parameter c_{ν_i}, which ranges between 0 and 0.5 and accounts for the chance of low-proficiency students guessing an item correctly:

$$P(x_{\nu_i} = 1) = c_{\nu_i} + (1 - c_{\nu_i})\frac{\exp(a_i(\theta_\nu - b_i))}{1 + \exp(a_i(\theta_\nu - b_i))}. \quad (3)$$

However, it is reasonable to acknowledge that a single physics testing item may encompass multiple concepts. Hence, student test results may not always conform to the assumptions of unidimensional IRT (i.e., a student can display multiple values for proficiency on one testing item). The compensatory multidimensional IRT model is commonly employed to address this concern by allowing for multidimensional proficiency and discrimination vectors for students $\boldsymbol{\theta}_\nu$ and items $\boldsymbol{a}_i$, while assuming a common difficulty parameter b_i for all concepts. The updated equation is:

$$P(x_{\nu_i} = 1) = c_{\nu_i} + (1 - c_{\nu_i})\frac{\exp(\boldsymbol{a}_i(\boldsymbol{\theta}_\nu - b_i))}{1 + \exp(\boldsymbol{a}_i(\boldsymbol{\theta}_\nu - b_i))}. \quad (4)$$

In this study, instead of assuming that the data perfectly fits a specific IRT model, we examined the model to identify deviations from these assumptions [14]. One crucial aspect is assessing whether the data are unidimensional. We employed the Bootstrap modified parallel analysis test (BMPAT) to examine the unidimensionality of the data [15, 16]. BMPAT is a statistical technique that evaluates the underlying dimensions of multiple-choice questions by examining the relationships between the dichotomously scored items.

B. Model Fit Indices

Several model fit statistical procedures and criteria are available to compare competing models. In this study, we employed the Akaike information criterion (AIC), Bayesian information criterion (BIC), and the adjusted version of the BIC

(SABIC) as the indices to evaluate how the candidate model fits our data. These indices enable comparison of model-data fit among a set of models using maximum-likelihood estimates of the parameters. Generally, smaller values of these indices mean better model fit. The AIC is calculated using the equation:

$$\mathrm{AIC} = -2\log L + 2p, \tag{5}$$

where L represents the likelihood function and p is the number of estimated parameters.

Similarly, the BIC is calculated using the equation:

$$\mathrm{BIC} = -2\log L + p\ln N. \tag{6}$$

In addition to the likelihood function L and the number of estimated parameters p, the BIC incorporates the sample size N in Eq. (6). The BIC is a more conservative estimate as it penalizes the selection of complex models [17]. The SABIC [18], on the other hand, uses N^* calculated by the equation:

$$N^* = \frac{N+2}{24} \tag{7}$$

instead of N. This adjustment has shown improved performance in studies involving a large number of parameters or small sample sizes [19, 20].

C. Data Collection

The data for this study were collected during the Fall 2022 and Spring 2023 semesters. The EMCS was administered along with a custom group work survey via Qualtrics to the IPLS course described in the background section. At the end of the survey, we added questions about the participants' demographics including gender, race/ethnicity, and disability status. Additionally, an attention check question was included among the items of the EMCS to allow us to identify if a participant was not reading the questions. Students completed the survey twice during the semester: once as a pretest in the first week of the semester and once as a posttest in the last week of the semester. Participation in the survey was anonymous, and at the end of the survey, students were requested to generate an identification code. The identification code contained five letters and four numbers from four pieces of information about the student and their major family events. The survey was announced as an extra credit assignment, and all students in the class were eligible to receive 0.5% of their course grade as extra credit if the response rate reached 80% or higher. A total of 454 students participated in the survey and passed the attention check, with 253 students completing the pretest and 201 students completing the posttest. 2 students failed the attention check and their data were removed from the data set being used in the present study.

D. Analysis

The statistical analysis in this study was conducted using R statistical software [21]. We utilized two R packages: `ltm` [22] for the unidimensional checks, and `mirt` [23] for the analyses related to multidimensional IRT.

The research question guiding this study was: What is the optimal statistical model for analyzing the structure of the test responses obtained from the EMCS? Considering the anticipated inclusion of more data in the coming academic years, the results and conclusions presented in the following sections are exploratory in nature.

III. RESULTS

A. Unidimensional Fit Analyses

To explore the dimensionality of the data, Bootstrap modified parallel analysis test (BMPAT) was conducted on the combined pretest and posttest results, as well as separately on the pretest results and posttest results. The unidimensional 1PL, 2PL, and 3PL models were tested, and the results of all nine tests are presented in Fig. 1. The vertical axes all show eigenvalue while the horizontal axes all show eigenvalue number. The black lines with circles represent eigenvalues calculated from the real data and the red lines with triangles represent eigenvalues calculated from the average simulated data. The average simulated data was generated by repeating the Monte Carlo simulations 100 times while assuming unidimensionality. The p-value was calculated for comparing the second eigenvalues of the real data and simulated data. The equation is:

$$p = \frac{n+1}{B+1}. \tag{8}$$

where n is the number of times when the simulated second eigenvalue is larger than the one from the real data and B is the number of the Monte Carlo simulations being repeated [22]. All tests indicated a statistically significant difference between the simulated and the real data with $p = 0.0099$.

Because $p = 0.0099$ in every test, we can reliably conclude that none of the simulated second eigenvalues are larger than the one from the real data. Therefore, the null hypothesis, that the data follow a unidimensional IRT model, can be rejected. This suggests that none of the tested models (combined, pretest, or posttest) adequately fit the data, indicating the presence of multidimensionality in the EMCS test items. The number of eigenvalues shown in Fig. 1 corresponds to the number of dimensions. The second eigenvalue values in the observed data were consistently higher than the simulated values across all model comparisons.

B. Multidimensionality fit Analysis

These findings prompted exploratory analyses to determine the optimal number of dimensions required for the model. A series of 3PL models ranging from 1 to 5 dimensions were examined, and the model fit statistics are presented in Table I. The AIC value decreased until the 4-dimensional model, while the BIC value decreased until the 2-dimensional model. The SABIC value, on the other hand, decreased until the 2-dimensional model.

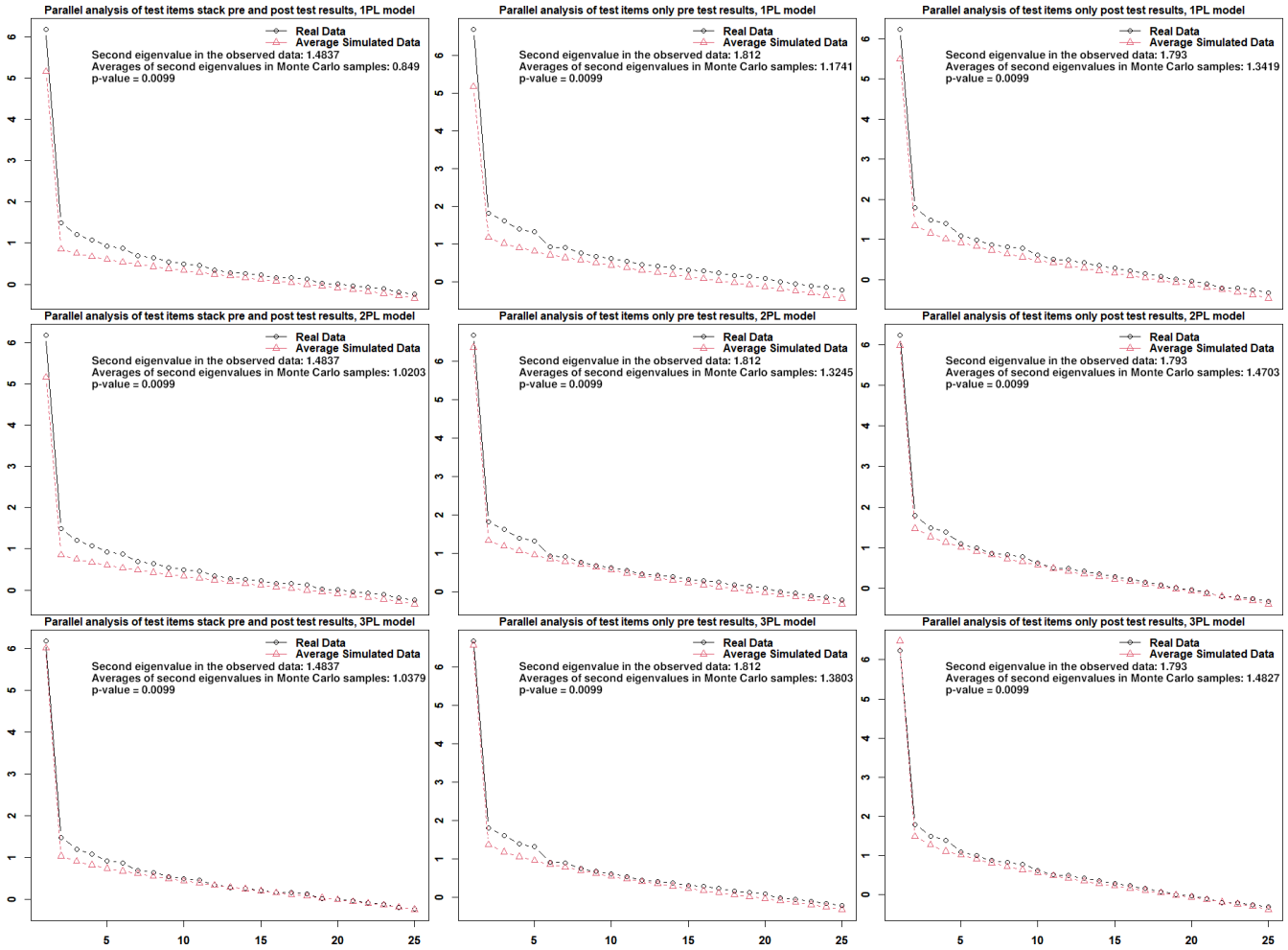

FIG. 1. Parallel analyses of the 1PL, 2PL, and 3PL models with stacking pre- and posttest data together and pre- or posttest data only. The vertical axes all show eigenvalue while the horizontal axes are all show eigenvalue number. The second eigenvalues in the observed data and simulations are listed, along with the corresponding p value.

Since SABIC and BIC tests are more conservative than AIC test, the optimal model may contain 2 or 3 dimensions. Considering the design of the EMCS, which assigned items 1, 2, 4, 6, 8, 9, 12, 13, 15, 17, 20, 22, 24, and 25 to energy and items 3, 5, 7, 10, 12, 14, 18, 19, 21, and 23 to momentum concepts with item 16 for both energy and momentum concepts, we explored three candidate models: two 2-dimensional 3PL models, where item 16 was assigned to either the energy or momentum group, and one 3-dimensional 3PL model, where item 16 formed a standalone third group. Since we would like to test whether the models with extra dimensionalities can align with the data better than the unidimensional model, another round of exploratory analyses was conducted to compare these candidate models against the 1-dimensional 3PL model as the baseline. The model fit statistics can be found in Table II. Since the 1-dimensional 3PL model has the smallest AIC, SABIC, and BIC values among all models listed in Table II, none of the candidate models demonstrated an advantage over the 1-dimensional 3PL model. The statistical evidence suggests that categorizing the EMCS items based on the designated testing concepts does not yield a better model compared to grouping all items together. Given the data of test responses reveals how students interpret the items on the test, our results showed students did not recognize the underlying concepts of items the same way as how the items were designed.

TABLE I. Multidimensional IRT model fit statistics with different dimensions

Dimensions	AIC	SABIC	BIC
1	13522.50	13593.34	13831.36
2	13401.31	13494.81	13809.00
3	13390.17	13505.39	13892.58
4	13352.11	13488.11	13945.12
5	13382.53	13538.36	14062.01

TABLE II. MIRT model fit statistics with different ways of grouping. "1D" is the 1-dimensional 3PL model. "2D16M" is the 2-dimensional 3PL model with assigning item 16 to the momentum group. "2D16E" is the 2-dimensional 3PL model with assigning item 16 to the energy group. "3D16ME" is the 3-dimensional 3PL model with item 16 being in a third group alone.

Model	AIC	SABIC	BIC
1D	13522.50	13593.34	13831.36
2D16M	13727.73	13798.56	14036.58
2D16E	13723.83	13794.67	14032.69
3D16ME	13749.74	13820.57	14058.59

IV. DISCUSSION

The dimensionality of conceptual inventory testing data is influenced by the interplay between how each item is designed and how students interpret and respond to those items. It is important to note that different student populations and institutions may yield varying results in this discussion. For instance, a study conducted by Wang and Bao [24] analyzed data from approximately two thousand students taking the FCI at the Ohio State University over a span of five years. In their study, both the pretest and posttest data aligned well with a unidimensional 3PL model. On the other hand, a study by Stewart et al., [25] involving over four thousand students from a different institution, suggested an optimal model with nine dimensions. Furthermore, other studies have suggested 5 or 6-dimensional models [26, 27].

In the present study, the analyses demonstrated the need for a multidimensional model to effectively explain the EMCS testing results. However, categorizing items based on the intentional testing concepts did not result in a better model compared to the unidimensional ones. This suggests that students may not interpret the items in the same way as intended by the test designers, and highlights the complexity of student understanding, as well as the need to further investigate the factors influencing item interpretation and response patterns.

When considering the use of multidimensional IRT, sample size becomes a crucial concern. Many of the previously cited works have employed samples with thousands of participants. However, according to the guidelines provided by Linacre, a sample size of around 150 participants can still achieve a high level of confidence, typically around 99% in most situations [28]. Given the exploratory nature of this study, the current sample size is deemed sufficient.

V. IMPLICATIONS

Based on the study and its conclusion, several implications can be suggested for researchers and educators. Firstly, it is important for researchers to carefully examine the structure of conceptual inventories and consider multidimensionality when analyzing and interpreting the results. This can provide a more accurate understanding of students' conceptual understanding and guide the development of more effective assessment tools.

Secondly, the findings emphasize the need for educators to be aware that students may interpret conceptual inventory items differently than intended. It is therefore critical to highlight the importance of aligning instruction and assessment to bridge any gaps in conceptual understanding and promote more accurate interpretations of the concepts being assessed.

Overall, this study underscores the ongoing need for rigorous examination of conceptual inventories and a deeper understanding of how students engage with conceptual inventory items. By incorporating these insights into research and practice, educators can better support students' conceptual learning and improve the effectiveness of conceptual assessments.

VI. FUTURE WORK

The present study opens up two potential directions for future research. Firstly, as we continue to accumulate data from more participants over multiple semesters, the MIRT analyses will gain increased statistical power, enabling more definitive conclusions regarding the structure of the testing items and how students comprehend the relevant concepts. With larger sample sizes, we will be better equipped to explore the multidimensionality of the EMCS test items and gain a deeper understanding of students' conceptual understanding.

Secondly, we plan to investigate how demographic factors influence student conceptual learning outcomes. Starting from the Fall 2022 semester, we have been collecting demographic information, including gender, ethnicity, and disability, potentially enabling us to examine the multifaceted fairness of the EMCS items and our IPLS course. By analyzing the data in conjunction with demographic factors, we can assess potential disparities in student performance. This research endeavor will provide insights into how demographic characteristics may interact with student understanding of energy and momentum concepts.

ACKNOWLEDGMENTS

This work is partially supported by the internal grant from University of Connecticut's College of Liberal Arts and Sciences.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, The Physics Teacher **30**, 141 (1992).

[2] S. Ramlo, The force and motion conceptual evaluation, Paper presented at the Annual Meeting of the Mid-Western Educational Research Association , 17 (2002).

[3] R. Chabay and B. Sherwood, Qualitative understanding and retention, AAPT Announcer **27**, 96 (1997).

[4] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, **69**, S12 (2001).

[5] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, American Journal of Physics **71**, 607 (2003).

[6] B. W. Junker, Some statistical models and computational methods that may be useful for cognitively-relevant assessment, Prepared for the National Research Council Committee on the Foundations of Assessment. Retrieved April **2**, 2001 (1999).

[7] G. J. Aubrecht and J. D. Aubrecht, Constructing objective tests, American Journal of Physics **51**, 613 (1983).

[8] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, Physical Review Special Topics - Physics Education Research **5** (2009).

[9] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, International Journal of Science Education **33**, 1289 (2010).

[10] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Physical Review Special Topics - Physics Education Research **10** (2014).

[11] R. J. Beichner and J. M. Saul, Introduction to the SCALE-UP (student-centered activities for large enrollment undergraduate programs) project, Proceedings of the International School of Physics (2003).

[12] Living Physics Portal (2023).

[13] R. Furr, *Psychometrics: An Introduction* (SAGE Publications, 2017).

[14] D. Andrich and I. Marais, *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*, Springer Texts in Education (Springer Nature Singapore, 2019).

[15] F. Drasgow and R. I. Lissak, Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses, Journal of Applied Psychology **68**, 363 (1983).

[16] H. Finch and P. Monahan, A bootstrap generalization of modified parallel analysis for irt dimensionality assessment, Applied Measurement in Education **21**, 119 (2008).

[17] S. I. Vrieze, Model selection and psychological theory: A discussion of the differences between the akaike information criterion (AIC) and the bayesian information criterion (BIC)., Psychological Methods **17**, 228 (2012).

[18] S. L. Sclove, Application of model-selection criteria to some problems in multivariate analysis, Psychometrika **52**, 333 (1987).

[19] J. M. Henson, S. P. Reise, and K. H. Kim, Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics, Structural Equation Modeling: A Multidisciplinary Journal **14**, 202 (2007).

[20] C.-C. Yang, Evaluating latent class analysis models in qualitative phenotype identification, Computational Statistics & Data Analysis **50**, 1090 (2006).

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2022).

[22] D. Rizopoulos, ltm: An R package for latent variable modeling and item response theory analyses, Journal of Statistical Software **17** (2006).

[23] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, Journal of Statistical Software **48** (2012).

[24] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, American Journal of Physics **78**, 1064 (2010).

[25] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the force concept inventory, Physical Review Physics Education Research **14** (2018).

[26] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a force concept inventory data set, Physical Review Special Topics - Physics Education Research **8** (2012).

[27] M. Semak, R. Dietz, R. Pearson, and C. Willis, Examining evolving performance on the force concept inventory using factor analysis, Physical Review Physics Education Research **13** (2017).

[28] J. Linacre, Sample size and item calibration stability, Rasch Measurement Transactions **7**, 328 (1994).