

Uncovering student conceptual structure by a multimethod evaluation of the energy and momentum conceptual survey

Xian Wu^{1,*}, Yaoguang Li^{2,3} and N. Sanjay Rebello^{4,5}

¹Department of Physics, *University of Connecticut*, Storrs, Connecticut 06269, USA

²Homer Babbidge Library, *University of Connecticut*, Storrs, Connecticut 06269, USA

³AGQ Solutions, South Windsor, Connecticut 06074, USA

⁴Department of Physics and Astronomy, *Purdue University*, West Lafayette, Indiana 47907, USA

⁵Department of Curriculum and Instruction, *Purdue University*, West Lafayette, Indiana 47907, USA



(Received 9 June 2025; accepted 10 October 2025; published 5 November 2025; corrected 12 December 2025)

The Energy and Momentum Conceptual Survey (EMCS) is an instrument designed to assess students' understanding of key physics concepts in introductory mechanics. In this study, we applied a combination of classical test theory (CTT), item response theory (IRT), and exploratory factor analysis (EFA) to a dataset of over 10 000 students collected across 5 semesters at a large U.S. university. CTT and IRT analyses identified several problematic items with weak discrimination, low reliability, or evidence of guessing. We then explored the factor structure of the full and reduced versions of the EMCS, comparing models with differing numbers of factors. While a five-factor model offered the best statistical fit, a four-factor model excluding three problematic items (Q16, Q22, and Q23) provided improved interpretability and conceptual coherence. In both models, energy-related items were separated into distinct clusters, suggesting multiple modes of student reasoning. Finally, we evaluated a theoretically grounded two-factor model reflecting the EMCS's design intent. This structure, representing energy and momentum as distinct yet stable factors, proved robust across model variations. The study demonstrated the best practice of exploratory factor analysis. Our findings demonstrate best practices for EFA while also highlighting the importance of instructional context in shaping learning gains. The low course-level gains observed in this study suggest the need to examine teaching practices along a continuum from traditional to active. Future work linking EMCS outcomes with detailed instructional-context data across multiple institutions will help clarify how different enactments of active teaching influence student conceptual understanding in physics.

DOI: [10.1103/kvph-l899](https://doi.org/10.1103/kvph-l899)

I. INTRODUCTION

A significant area of focus in physics education research is the development and use of research-based assessments (RBAs). Over the years, numerous RBAs have been created to evaluate students' conceptual understanding, mathematics, scientific reasoning, and attitude in physics. PhysPort, a website that provides resources for physics educators, currently lists 118 RBAs across various topics [1]. These assessments serve multiple academic purposes: physics instructors use them to measure the impact of their courses, curriculum developers rely on them to evaluate new or revised instructional materials, and education researchers analyze RBA data to explore patterns in students' ideas.

The development and validation of an RBA is a rigorous and iterative process. It typically begins with interviews and/or a review of existing literature to gather insights into students' thinking, prior knowledge, and common difficulties. Based on both student-derived information and expert understanding of the domain, survey or test items are constructed to probe key conceptual, procedural, or attitudinal dimensions. These items are then refined through multiple rounds of testing with student populations to evaluate clarity, reliability, and validity. Revisions are made as needed, often through iterative cycles of statistical analysis and expert review, before the final version of the RBA is published and broadly adopted [2–5]. Nevertheless, validating and revisiting a published RBA remains important, since its reliability and applicability may shift across different contexts, such as institutions, countries, or even successive student generations. This continuing process ensures that an RBA remains both psychometrically sound and educationally meaningful over time.

Beyond their role in validation, statistical analysis of RBA data has become increasingly prominent in physics

*Contact author: xian.wu@uconn.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

education research. Ding and Beichner [6] provided an overview of commonly used approaches for analyzing such data, including methods for evaluating item quality and examining student learning outcomes. More recent studies have incorporated broader techniques such as network analysis [7], multidimensional item response theory [8–11], and regression modeling [12]. Within introductory mechanics, several RBAs that take the form of conceptual inventories have been developed and extensively studied. The Force Concept Inventory (FCI) remains one of the most widely studied assessments [13]. It has been used extensively to analyze students' conceptions [14–16], conceptual structures [17–22], and issues related to fairness across gender and race [23–25]. Other studies have examined how students interpret different representations used in questions [26,27]. The Force and Motion Conceptual Evaluation (FMCE) [28] is another commonly used inventory to be compared with FCI [29–31] due to the similar concept coverage.

Relatively little research has focused on the Energy and Momentum Conceptual Survey (EMCS) [32], despite the foundational importance of work, energy, impulse, and momentum in introductory mechanics courses [33,34]. The EMCS is a 25-item multiple-choice assessment, with fourteen items targeting energy, ten items targeting momentum, and one item integrating both [32]. The full instrument is available through the PhysPort website [35]. The concepts covered by the EMCS are not only central to introductory mechanics but are also known to be persistently challenging for students to master, often involving abstract reasoning, multistep problem solving, and the coordination of multiple representations [36,37]. Therefore, assessing and improving conceptual understanding in these areas is critical for supporting student success in physics and related disciplines.

One notable advantage of using the EMCS for statistical analysis is that it focuses solely on two major concepts: energy and momentum, allowing for a clearer examination of how statistical models align with both student and expert conceptual frameworks, in contrast to other inventories that span a broader range of mechanical topics. Wu *et al.* [11] analyzed the dimensionality of the EMCS data collected from a small ($N = 454$) student population and found more than one existing dimension. However, a thorough study with more students is needed to obtain a definitive result.

The present study addresses this need while also showcasing the value of a mixed-method psychometric approach. Inspired by recent advances in assessment methodology, we integrate classical test theory (CTT), item response theory (IRT), and exploratory factor analysis (EFA) to evaluate the EMCS in a comprehensive research-grounded manner. This combination of methods has been proposed in frameworks for instrument validation [38] but has rarely been applied together in physics education research, particularly to a conceptual inventory with such

a focused theoretical design. By pairing a rigorous multi-method analysis with the unique structure of the EMCS, this study provides both methodological insights into best practices for RBA validation and conceptual insights into how students reason about fundamental ideas of energy and momentum. Recent work in statistical psychology has further emphasized the importance of following best practices in factor analysis [39–42], and our study contributes to this effort by applying those recommendations systematically to a large dataset.

Therefore, the primary goal of this study is to apply validated practices in factor analysis to EMCS data in order to examine the relationship between statistically derived models based on real-world student data and expert-conceived conceptual structures. We aim to understand how students' conceptual organization corresponds with the assessment's intended design and how this alignment is influenced by the psychometric quality of individual items.

To achieve this goal, we address the following research questions:

RQ1: Which items on the EMCS exhibit weak psychometric properties, as identified through comprehensive analyses using CTT and IRT?

RQ2: How does the conceptual structure derived from statistical analyses compare with the expert-defined conceptual structure?

II. THEORETICAL BACKGROUND

A. Energy and momentum conceptual survey

The Energy and Momentum Conceptual Survey (EMCS) is a multiple-choice assessment consisting of 25 items, each with 5 answer options. It is designed to evaluate students' conceptual understanding of key energy and momentum topics typically taught in introductory physics courses. Specifically, the EMCS targets the work-energy theorem, conservation of mechanical energy, impulse-momentum theorem, and conservation of momentum.

The development and validation of the EMCS followed a multistage process involving free-response diagnostics, student interviews, expert reviews, and field testing. Prior to its publication, the instrument was administered to over 3000 students across more than 30 courses over multiple semesters, providing a robust dataset for validation [32].

The test developers categorized the items into two primary conceptual domains: momentum-related and energy-related. Fourteen items (Questions 1, 2, 4, 6, 8, 9, 12, 13, 15, 17, 20, 22, 24, and 25) target energy concepts, while ten items (Questions 3, 5, 7, 10, 11, 14, 18, 19, 21, and 23) focus on momentum. Question 16, which features a ballistic pendulum, integrates both energy and momentum concepts (For consistency, in the remainder of this paper, we refer to items as Q01–Q25.). Reported average pre-instruction scores are around 30%, increasing to approximately 50% postinstruction, with a typical normalized gain

of 0.29. Notably, Q16 has consistently shown a negative learning gain, highlighting students' persistent difficulty in integrating energy and momentum concepts.

An initial factor analysis conducted by the test developers revealed no dominant factor, with the largest single factor explaining less than 15% of the variance. This result suggests that multiple underlying factors may be necessary to account for student performance on the EMCS.

More recently, a study examined EMCS performance among both introductory-level engineering students and upper-level physics majors [43]. The study included over 300 introductory students and about 50 physics majors. Q03, Q05, Q06, Q08, Q09, Q10, Q12, Q13, Q16, Q17, Q22, Q23, and Q24 were identified as the most challenging for both groups. Consistent with previous findings, introductory students again showed a negative gain on Item 16. In contrast, physics majors demonstrated moderate improvement on that item (normalized gain $g = 0.32$). This recurring difficulty among less experienced students underscores the conceptual complexity of items requiring the integration of energy and momentum ideas.

B. Classical test theory

Classical test theory (CTT) [44] is a foundational framework widely used in the validation of conceptual inventories. It helps researchers assess whether an instrument meaningfully measures the factors it is intended to evaluate. CTT operates under the assumption that a student's observed score on a test is composed of their true score plus some random error. CTT-based analyses focus on several key metrics: item difficulty, item discrimination, point-biserial correlation, and Cronbach's alpha.

Item difficulty is defined as the proportion of students who answered an item correctly out of those who attempted it. The value ranges from 0 to 1, with higher values indicating easier items for the tested population. For the EMCS items with five options, random guessing yields a 20% success rate. Therefore, in this study, items with a difficulty value below 0.20 are considered excessively difficult. [45].

Item discrimination measures how well an item distinguishes between students with higher and lower overall test performance. It is typically calculated as the difference in item difficulty between the top-performing group and the bottom-performing group, often defined as the top and bottom 25% or 27% of scores [6,45]. In the current study, we used the 27% cutoff [46]. Discrimination values range from -1 to 1 , with values above 0.40 generally considered very good, and values below 0.20 suggesting poor discrimination and the potential need for item revision [47].

The point-biserial correlation evaluates whether students who score higher overall are more likely to answer a particular item correctly. It is a measure of how well an item correlates with total test performance, with values also ranging from -1 to 1 . Higher values indicate stronger

alignment with the overall test. A point-biserial value above 0.20 is typically considered acceptable [48].

Cronbach's alpha provides an estimate of test reliability, defined as the consistency of scores across repeated administrations of the test. A higher alpha suggests that the test produces stable scores across administrations. Values above 0.70 are generally accepted as indicative of acceptable reliability [49].

CTT-based statistics are commonly used as an initial step in test validation. They are often followed by factor analysis, which explores the underlying structure of the assessment [38,44]. Items may cluster together statistically when they share conceptual or linguistic features, revealing potential subdimensions within the test.

C. Item response theory

Unlike CTT, item response theory (IRT) [50] uses probabilistic models to represent how likely a student is to answer an item correctly based on their underlying ability. These models are constructed to directly model student responses to individual test items, making them especially valuable for analyzing performance on conceptual inventories.

One progression of IRT models is based on the number of item-specific parameters estimated. The one-parameter logistic (1PL) model estimates only item difficulty. The two-parameter logistic (2PL) model includes both item difficulty and item discrimination. The three-parameter logistic (3PL) model adds a guessing parameter as the third component. Finally, the four-parameter logistic (4PL) model includes an additional inattention or slipping parameter. This study uses the 3PL model, which balances complexity and interpretability, particularly for assessments like the EMCS that include multiple-choice items with a non-negligible probability of guessing.

The 3PL item response function (IRF) is defined as

$$P_j(\theta_i) = c_j + (1 - c_j) \cdot \frac{1}{1 + e^{-a_j(\theta_i - b_j)}}$$

In this function, θ_i represents the latent ability of the student i , and $P_j(\theta_i)$ is the probability that the student answers the item j correctly. The parameter a_j denotes item discrimination, which controls the steepness of the IRF. Higher values of a_j indicate that the item more effectively distinguishes between students of differing abilities. Typical values for a_j range from 0.5 to 2.5 ; values lower than 0.5 are generally considered not sensitive to differences in ability. The difficulty parameter b_j represents the ability level at which a student has a 50% chance of answering the item correctly above the guessing level. Values of b_j commonly range between -3 and $+3$. The parameter c_j is the guessing factor, which models the lower bound of the IRF. It reflects the probability of answering the item correctly by chance. For five-option multiple-choice

questions such as those on the EMCS, a typical value for c_j is approximately 0.20.

When the 3PL model is simplified to the 1PL model, the guessing parameter c_j for each item is set to zero, and the discrimination parameter a_j is set to one. Only the difficulty parameter b_j is allowed to vary. The inclusion of the guessing parameter in the 3PL model allows for a more realistic representation of student performance on multiple-choice assessments like the EMCS, where lower-ability students may still answer correctly by guessing.

D. Overview of exploratory factor analysis

Factor analysis comprises a suite of statistical techniques used to investigate the underlying structure among a set of observed variables, typically referred to as factors [39,51]. The central goal of factor analysis is to identify a model with the fewest number of latent factors that can explain a substantial portion of the variance in the dataset.

There are two primary types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is generally used when there is no prior expectation about the factor structure. In contrast, CFA is used to test how well a theoretical or hypothesized factor model fits the observed data. Due to their different purposes and underlying algorithms, it is common for models suggested by EFA and CFA to diverge, particularly in cases involving complex factor structures [21,22]. The results of both approaches can also be influenced by practical factors, such as the specific statistical techniques used, the characteristics of the student population, and the conditions under which the assessment is administered. In this study, we focus on EFA, guided by recommendations from recent reviews by statistical psychologists [39–42].

Before applying any factor analyses, it is necessary to assess whether the data meet the assumptions required for valid analysis. Two commonly used diagnostics for EFA are Bartlett’s test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy [52–54]. Both tests evaluate whether sufficient correlations exist among the variables to justify the application of factor analysis.

Bartlett’s test of sphericity examines whether the correlation matrix significantly differs from an identity matrix. It is calculated using the formula:

$$\chi^2 = -\left(N - 1 - \frac{2p + 5}{6}\right) \cdot \ln |R|$$

where N is the sample size, p is the number of variables (i.e., test items), and $|R|$ is the determinant of the correlation matrix. A significant result indicates that the correlations among variables are adequate for factor analysis.

The KMO measure of sampling adequacy assesses whether the partial correlations among variables are small, suggesting that latent factors are likely present. KMO values range from 0.00 to 1.00, with values above 0.60

considered acceptable for proceeding with EFA. If the KMO value is low, it means that a large proportion of the variance of the data cannot be explained by underlying factors. It is recommended to take remedial actions to improve the dataset’s suitability for factor analysis.

The distinction between principal component analysis (PCA) and EFA has been a topic of long-standing debate among statisticians. From a methodological standpoint, PCA and EFA are often viewed as distinct types of analysis. However, a practice-based perspective sometimes blurs this distinction, as the two methods can yield similar results [55]. PCA is frequently used in practice, especially because it is the default extraction method in software such as SPSS [56].

In this study, we follow Howard’s recommendation [40] and treat PCA as “not a true form of factor analysis” (p. 52). Although PCA may produce similar results to EFA, differences can arise, particularly in terms of underlying assumptions and interpretation. Based on the guidance of Conway and Huffcutt [57], we use principal axis factoring (PAF), a method of common factor extraction, as the technique for conducting EFA in this study.

III. METHODS

The participants in this study were introductory-level physics students from a land-grant public research university located in the Midwest region of the United States. Data were collected over five consecutive semesters, from Spring 2021 to Spring 2023. During this period, students returned to in-person instruction following the COVID-19 pandemic. In total, 6343 students completed the EMCS as a pretest before instruction, and 5996 students completed the EMCS as a post-test after instruction. The dataset contains 529 unanswered item cells and 12 students who submitted all incorrect responses. These 12 students were excluded from all subsequent analyses due to their lack of valid response patterns. All statistical analyses were conducted using R version 4.4.2 [58]. The DPLYR [59] and PURRR [60] packages were used for structure and data processing. The PSYCH package [61] was used to conduct CTT analyses and EFA, while the MIRT package [62] was employed for IRT modeling. The QGRAPH package [63] was used to plot the network diagrams based on the EFA results.

IV. RESULTS

A. Baseline results

Descriptive statistics for all items are presented in Table I. The average pretest score across all students was 32.56% (SD = 45.94%), and the average post-test score was 34.34% (SD = 46.80%). Item-level learning gains were calculated using the normalized gain formula:

$$\text{Gain} = \frac{\text{Post} - \text{Pre}}{1 - \text{Pre}}$$

TABLE I. Descriptive statistics and item-level psychometric properties of the Energy and Momentum Conceptual Survey (EMCS). Items are labeled Q1–Q25, with each categorized by the test developers as “E” (energy-related), “M” (momentum-related), or “E&M” (both energy and momentum). “Pre-test” and “Post-test” represent the average correctness rate ($\pm$ standard deviation) before and after instruction, respectively. “Gain” is the average normalized learning gain per item. “CTT_Diff” and “IRT_Diff” indicate item difficulty values calculated via Classical Test Theory (CTT) and 3PL Item Response Theory (IRT), respectively. “CTT_Disc” and “IRT_Disc” represent item discrimination indices from CTT and IRT. “Bisr” refers to the point-biserial correlation (CTT). “IRT_Gue” is the guessing parameter estimated via IRT. “ α_{if_rm} ” is Cronbach’s alpha for the overall test if the item is removed. Summary rows report the average statistics for energy-related items [Average (E)], momentum-related items [Average (M)], and all items [Average (All)]. Values not meeting commonly accepted psychometric thresholds are highlighted in bold. NA stands for not applicable.

Item	Pre-test	Post-test	Gain	CTT_Diff	IRT_Diff	CTT_Disc	IRT_Disc	Bisr	IRT_Gue	α_{if_rm}
Q01 (E)	0.2856 $\pm$ 0.4518	0.4583 $\pm$ 0.4983	0.24	0.3694	1.0386	0.4283	1.8404	0.3336	0.1860	0.7458
Q02 (E)	0.4118 $\pm$ 0.4922	0.4629 $\pm$ 0.4987	0.09	0.4366	0.2477	0.5103	1.4336	0.4206	0.0125	0.7398
Q03 (M)	0.4212 $\pm$ 0.4938	0.3961 $\pm$ 0.4891	−0.06	0.4087	1.0763	0.3111	0.6516	0.2065	0.1032	0.7540
Q04 (E)	0.4428 $\pm$ 0.4968	0.4173 $\pm$ 0.4931	−0.06	0.4302	0.6567	0.4493	1.5053	0.3538	0.1549	0.7443
Q05 (M)	0.2161 $\pm$ 0.4116	0.2672 $\pm$ 0.4426	0.07	0.2408	1.6048	0.3723	2.0343	0.2852	0.1465	0.7489
Q06 (E)	0.2187 $\pm$ 0.4134	0.3647 $\pm$ 0.4814	0.19	0.2896	1.5596	0.3530	2.3626	0.2591	0.2068	0.7504
Q07 (M)	0.4888 $\pm$ 0.4999	0.4530 $\pm$ 0.4978	−0.07	0.4713	0.0734	0.5572	1.9354	0.4723	0.0047	0.7362
Q08 (E)	0.2730 $\pm$ 0.4456	0.3387 $\pm$ 0.4733	0.09	0.3048	1.3047	0.4082	1.8293	0.3166	0.1634	0.7469
Q09 (E)	0.3029 $\pm$ 0.4595	0.2427 $\pm$ 0.4287	−0.20	0.2735	1.8638	0.3141	1.3025	0.2198	0.1681	0.7527
Q10 (M)	0.2638 $\pm$ 0.4407	0.3054 $\pm$ 0.4606	0.06	0.2840	1.5585	0.3573	2.1580	0.2643	0.1934	0.7501
Q11 (M)	0.3020 $\pm$ 0.4591	0.2671 $\pm$ 0.4425	−0.12	0.2848	0.9514	0.5257	1.8960	0.4463	0.0643	0.7390
Q12 (E)	0.2539 $\pm$ 0.4353	0.2993 $\pm$ 0.4580	0.06	0.2757	1.7033	0.3224	2.3877	0.2282	0.2095	0.7522
Q13 (E)	0.4183 $\pm$ 0.4933	0.3897 $\pm$ 0.4877	−0.07	0.3940	1.4681	0.3340	1.2424	0.2313	0.2631	0.7524
Q14 (M)	0.4631 $\pm$ 0.4987	0.3727 $\pm$ 0.4836	−0.17	0.4179	1.1763	0.3759	1.6689	0.2749	0.2766	0.7496
Q15 (E)	0.3469 $\pm$ 0.4760	0.3017 $\pm$ 0.4590	−0.13	0.3246	1.0066	0.4752	1.8510	0.3876	0.1261	0.7424
Q16 (E&M)	0.3645 $\pm$ 0.4813	0.3924 $\pm$ 0.4883	0.04	0.3778	2.0121	0.2345	0.3629	0.1280	0.0776	0.7589
Q17 (E)	0.2890 $\pm$ 0.4533	0.3264 $\pm$ 0.4689	0.05	0.3070	1.8270	0.2910	2.1832	0.1920	0.2509	0.7545
Q18 (M)	0.2965 $\pm$ 0.4568	0.3447 $\pm$ 0.4753	0.07	0.3197	0.9555	0.4442	1.1923	0.3542	0.0503	0.7445
Q19 (M)	0.4454 $\pm$ 0.4971	0.4269 $\pm$ 0.4947	−0.04	0.4363	0.6353	0.4380	1.3570	0.3413	0.1459	0.7452
Q20 (E)	0.3081 $\pm$ 0.4617	0.3706 $\pm$ 0.4830	0.09	0.3384	0.9397	0.5021	2.7194	0.4161	0.1608	0.7405
Q21 (M)	0.4021 $\pm$ 0.4904	0.4059 $\pm$ 0.4911	0.01	0.4036	1.1645	0.3909	1.9005	0.2916	0.2646	0.7485
Q22 (E)	0.1962 $\pm$ 0.3972	0.2004 $\pm$ 0.4003	0.01	0.1981	2.7276	0.1958	1.5205	0.1076	0.1642	0.7583
Q23 (M)	0.2959 $\pm$ 0.4565	0.3232 $\pm$ 0.4677	0.04	0.3090	1.8322	0.2353	4.2727	0.1339	0.2786	0.7580
Q24 (E)	0.2109 $\pm$ 0.4080	0.2697 $\pm$ 0.4439	0.07	0.2393	1.7444	0.3227	2.4102	0.2329	0.1753	0.7519
Q25 (E)	0.2219 $\pm$ 0.4156	0.1890 $\pm$ 0.3915	−0.15	0.2056	1.7682	0.3617	1.4111	0.2789	0.0881	0.7493
Average (E)	0.2985 $\pm$ 0.4576	0.3307 $\pm$ 0.4705	0.02	0.3133	1.4182	0.3763	1.8571	NA	0.1664	NA
Average (M)	0.3595 $\pm$ 0.4799	0.3562 $\pm$ 0.4789	−0.02	0.3576	1.1028	0.4008	1.9067	NA	0.1528	NA
Average (All)	0.3256 $\pm$ 0.4594	0.3434 $\pm$ 0.4680	0.00	0.3363	1.3159	0.3804	1.8172	NA	0.1574	NA

where “Post” and “Pre” refer to the average item correctness rates on the post-test and pretest, respectively [64]. Item scores were scaled from 0 to 1 prior to this calculation. For items where the post-test average was lower than the pretest average, the following alternative formula was used [65]:

$$\text{Gain} = \frac{\text{Post} - \text{Pre}}{\text{Pre}}$$

Except for Items 1 and 6, most items demonstrated moderate, negligible, or even negative learning gains. We acknowledge that various methods exist for calculating learning gains, and the validity of different approaches remains a topic of debate in the literature [66,67]. Evaluating those methods is beyond the scope of this study.

In terms of overall performance, the pretest results are comparable to those reported in earlier studies. For

example, Singh and Rosengrant [32] reported pretest averages ranging from 24% to 36% and a post-test average of 49%. Our dataset yielded a pretest average of 32.6%, which falls within that range. However, the post-test average of 34.3% in our study is substantially lower than the 49% reported in the original study, suggesting limited learning gains across the student population. Notably, no item in the current dataset showed a post-test correctness rate exceeding 50%.

B. CTT and IRT results

The results of the CTT and IRT analyses are presented in Table I. From the CTT framework, three primary metrics were extracted: CTT difficulty, CTT discrimination, and point-biserial correlation. Additionally, Cronbach’s alpha was computed for the entire test, as well as for the test when each item was individually removed.

As previously discussed, CTT difficulty values range from 0 to 1, with higher values indicating easier items. Values below 0.20 suggest that an item may be excessively difficult. Most items in the dataset fall within the acceptable range, except Q22, which was considerably difficult. CTT discrimination values range from -1 to 1 , with values below 0.2 typically considered too low. Again, Q22 was the only one failing to meet this criterion. Regarding point-biserial correlation, values also range from -1 to 1 , with higher values indicating better alignment between item performance and total test scores. Values below 0.2 are generally viewed as problematic. Q16, Q17, Q22, and Q23 fell below this threshold.

The Cronbach's alpha for the full test was 0.7563 , which is considered acceptable. However, when examining alpha if an item is removed, the removal of Q16, Q22, and Q23 led to higher reliability estimates, suggesting that these items may detract from the overall coherence of the test.

To gain a deeper understanding of the test's psychometric structure, we also conducted IRT analyses. In the IRT framework, difficulty values typically range from -3 to $+3$, with higher values indicating more difficult items. Q22, with an IRT difficulty value of 2.73 , is indicated to be much harder than other questions in the test, with the second-highest IRT difficulty value of 2.01 . Since there was no negative value for IRT difficulty, no item was found to be too easy for students. IRT discrimination values are ideally between 0.5 and 2.5 . Q16, with a discrimination value of 0.36 , did not meet this threshold, indicating poor differentiation between high- and low-ability students. Conversely, Q23 exhibited a high IRT discrimination value of 4.27 , while its CTT discrimination was relatively low (0.235). Its point-biserial correlation was also below the commonly accepted threshold of 0.2 , with a value of 0.134 . Since the point-biserial is an alternative indicator of an item's ability to differentiate among students [45], this discrepancy raises further concerns about the item's validity. Taken together, these findings suggest that Item 23 may be problematic in accurately assessing student ability.

The guessing parameter for a five-option multiple-choice test is typically around 0.20 . However, Q13, Q14, Q17, Q21, and Q23 had much higher guessing values, suggesting a higher-than-expected rate of random guessing on those items.

All parameter values that failed to meet commonly accepted thresholds are highlighted in Table I. While a direct comparison of CTT and IRT results is beyond the scope of this study, conducting both sets of analyses allows for a more comprehensive evaluation of the instrument. Overall, Q16, Q22, and Q23 were identified as the most problematic across both frameworks. In subsequent dimensionality analyses, we consider the removal of these items to explore whether doing so results in a clearer factor structure, such as fewer underlying dimensions or weaker interfactor correlations.

C. EFA results

To assess the suitability of the dataset for EFA, Bartlett's test of sphericity and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy were conducted. Bartlett's test yielded a chi-square value of $24\,266.94$ with 300 degrees of freedom and a p value less than $.001$, indicating that the correlation matrix significantly differs from an identity matrix and that sufficient correlations exist among items to justify factor analysis. The overall KMO value was 0.90 , which is considered “marvelous” according to Kaiser's classification [68], suggesting that the shared variance among items is suitable for factor extraction. The item-level measures of sampling adequacy ranged from 0.78 to 0.91 , all exceeding the recommended threshold of 0.70 . These values confirm that each item contributes meaningfully to a common factor structure and that the data are appropriate for subsequent EFA.

To determine the optimal number of factors to retain, Horn's parallel analysis was conducted using principal axis factoring (PAF) as the extraction method with 100 random permutations. The analysis compared the actual eigenvalues from the dataset with those generated from randomly simulated datasets of the same size. The results indicated that the first nine observed eigenvalues exceeded the corresponding random values, suggesting the retention of up to nine factors (see Fig. 4 in the Appendix). However, as parallel analysis typically provides an upper bound rather than a definitive factor solution, further model comparison was necessary to evaluate more parsimonious alternatives.

We examined model fit statistics for one- through seven-factor solutions using PAF, assessing root mean square error of approximation (RMSEA), Tucker–Lewis index (TLI), and Bayesian information criterion (BIC). While both RMSEA and TLI improved consistently with the addition of more factors, the BIC reached its minimum at the five-factor model ($BIC = -1025.57$) and increased thereafter, suggesting diminishing returns and potential overfitting beyond five factors (Table II). A scree plot [Fig. 1(a)] was also generated using the eigenvalues from the 25-item dataset, showing a steep decline between the first and second factors, followed by a gradual tapering. The first four eigenvalues (3.917 , 1.237 , 1.122 , and 1.051) were above 1.0 , with the fifth value (1.003) just above the Kaiser criterion cutoff. Subsequent eigenvalues dropped below 1.0 , indicating that additional factors explained limited variance. Together, these results suggest that while nine factors may statistically account for shared variance, a five-factor solution provides a more interpretable and parsimonious structure.

To explore whether a cleaner factor structure could be achieved by removing problematic items, we repeated the EFA after excluding Q16, Q22, and Q23, which were identified in prior CTT and IRT analyses as having poor psychometric properties. This reduced the item pool from 25 to 22 items. Model fit statistics for one- through

TABLE II. Results of factor analysis based on eigenvalue extraction for one- through seven-factor models, comparing analyses using all 25 items (25Its) versus a reduced set of 22 items (22Its; excluding Items 16, 22, and 23). For each model, we report the eigenvalue of the added factor, RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index; BIC, Bayesian information criterion, and the cumulative percentage of variance explained (VarExp) by the corresponding factor model. Note that “VarExp” represents the proportion of total test variance explained by a model with n factors (e.g., a two-factor model explains 14% of the variance), not the variance explained by the n th factor.

Number of factors	Eigenvalue (25Its)	RMSEA (25Its)	TLI (25Its)	BIC (25Its)	VarExp (25Its) (%)	Eigenvalue (22Its)	RMSEA (22Its)	TLI (22Its)	BIC (22Its)	VarExp (22Its) (%)
1	3.917	0.0263	0.897	-51.0627	12	3.862	0.0271	0.913	71.6628	14
2	1.237	0.0221	0.928	-652.759	14	1.189	0.0224	0.941	-463.758	15
3	1.122	0.0188	0.947	-953.256	15	1.053	0.0191	0.957	-678.744	17
4	1.051	0.0171	0.957	-1015.98	16	1.020	0.0171	0.966	-734.990	18
5	1.003	0.0155	0.965	-1025.57	17	0.970	0.0154	0.972	-728.610	19
6	0.987	0.0139	0.971	-1004.35	18	0.957	0.0129	0.980	-731.794	19
7	0.977	0.0126	0.977	-950.896	18	0.942	0.0120	0.982	-653.582	20

seven-factor solutions were reevaluated using the same PAF method. The BIC for the reduced set reached its minimum at the four-factor model (BIC = -734.99) before

increasing, suggesting that the reduced dataset may support an even more parsimonious structure than the original. The RMSEA and TLI values for the four-factor model were 0.0171 and 0.966, respectively, both indicating strong model fit (Table II).

This was further supported by eigenvalue analysis and visual inspection of the revised scree plot [Fig. 1(b)]. The first four eigenvalues (3.862, 1.189, 1.053, and 1.020) were above the 1.0 threshold, while the fifth (0.970) fell just below it, suggesting that the key inflection point in variance explained occurs after the fourth factor. This pattern aligns with the Kaiser criterion and confirms the relative strength of the four-factor model.

In summary, removing Q16, Q22, and Q23 improved the factor structure, led to better overall model fit, and supported a more concise and interpretable four-factor solution. This refined model not only aligns more closely with theoretical expectations but also enhances the psychometric robustness and clarity of the EMCS instrument.

V. DISCUSSION

A. Conceptual analysis of problematic items (Q16, Q22, and Q23)

Guided by the CTT and IRT results, items Q16, Q22, and Q23 emerged as the most problematic questions on the EMCS with the current dataset. Below, we consider likely conceptual reasons.

Q16 is the only item that requires coordinating both energy and momentum concepts to arrive at the correct answer [32]. Its weak alignment with other EMCS items (low point-biserial and a reduction in Cronbach’s α when retained) is therefore unsurprising. Importantly, Q16 is not among the hardest items, and its guessing parameter is relatively low, indicating that it functions as a reasonable physics question but does not match the EMCS.

Q22 is among the most difficult items, whereas Q23 is of moderate difficulty; both nonetheless show weak alignment with all the other EMCS items. A common feature is the

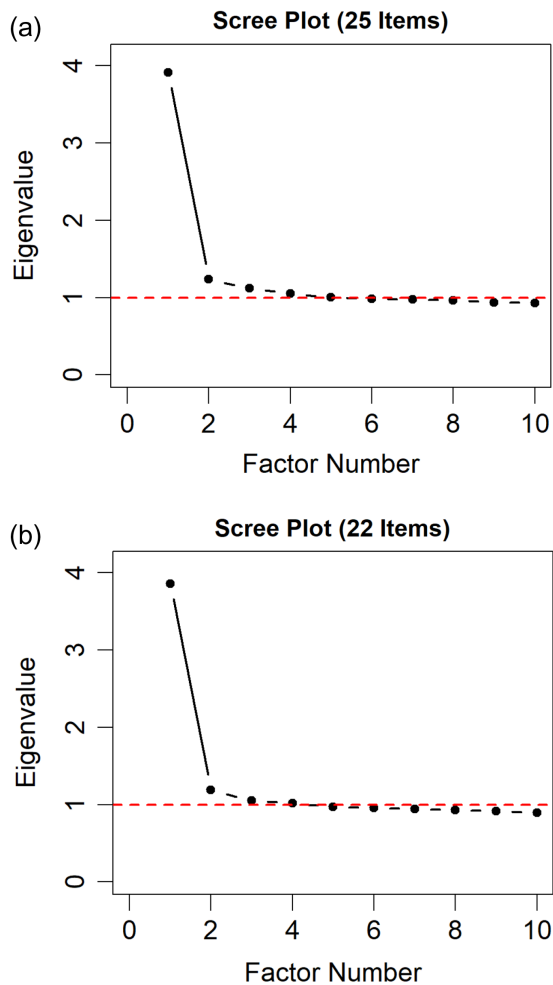


FIG. 1. (a) Scree plot of eigenvalues using all 25 items (b) Scree plot of eigenvalues using 22 items (excluding items 16, 22, and 23).

absence of explicit disciplinary cueing: unlike most EMCS items, the stems for Q22 and Q23 do not include the words *energy*, *work*, or *momentum*. This absence likely impedes activation of the relevant knowledge schema and selection of the intended solution strategy. Prior research suggests that students often default to kinematic/force-based reasoning for work or energy questions [69]. Thus, even when an item targets an energy or momentum construct, failure to give the signal clearly in the wording can encourage an inappropriate problem-solving frame and depress discrimination.

B. Statistical result-informed models

Interpretations of EFA results are often complicated by their dependence on the specific population studied and the potential instability of multifactor models. Because of this, we take a cautious and focused approach in our discussion. Rather than speculating on the placement of individual items, such as why a momentum-related question might load with energy-related ones, we center our attention on the larger structural patterns that emerge from the analysis. In particular, we examine the composition and coherence of major factors, especially those that appear to represent interpretable clusters of conceptual understanding. This approach also allows us to assess how statistical models derived from student response data align—or diverge from—expert-defined conceptual structures.

Our interpretation begins with two models that offered the best statistical fit: a five-factor solution including all 25 EMCS items, and a four-factor solution generated after removing Items 16, 22, and 23, identified as psychometrically problematic through CTT and IRT analyses. These two models were selected not only because they yielded the strongest model fit indices (see Table II) but also because they captured the richest patterns in the data while maintaining interpretability.

Before delving into factor composition, it is important to recall that the input to these analyses was binary correctness data. Student responses thus reflect a compound of conceptual understanding and interpretive behavior, shaped by both their grasp of physics and their perception of item structure. While conceptual inventories are built to probe student understanding, students often approach items with different problem-solving strategies than instructors intend, sometimes relying on surface features or familiar formulas instead of underlying concepts [70]. This complexity is precisely what research-based assessments like the EMCS attempt to navigate, combining expert-informed content with empirically grounded item construction to infer student conceptions.

As we can see in Fig. 2, in the five-factor model, a dominant cluster (factor 1) contains a core of momentum-related items, including seven of the ten items originally categorized as momentum-based, along with one energy-related item and item 16, which blends energy and

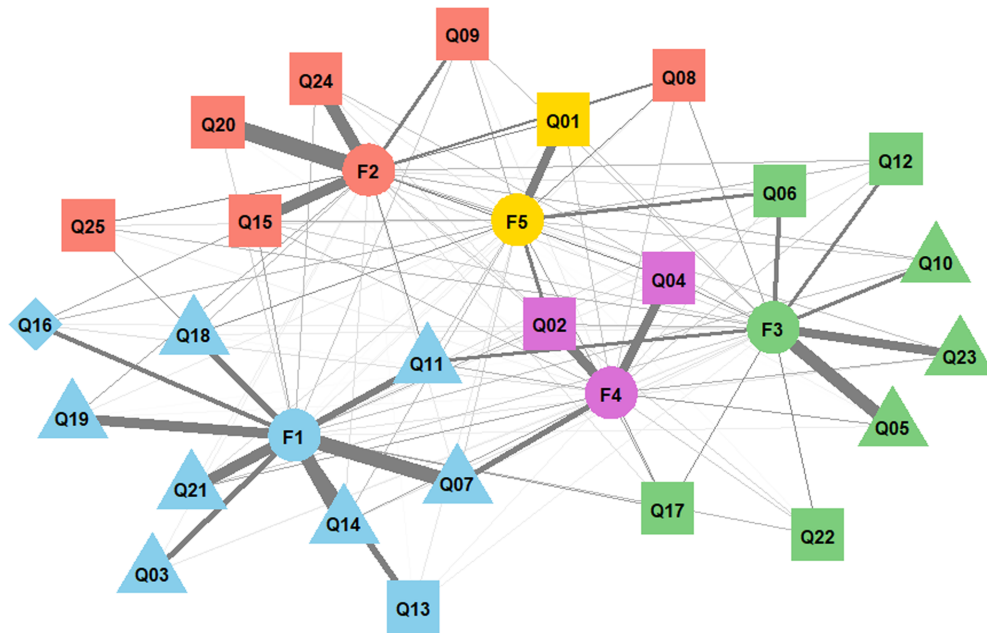
momentum concepts. Three other factors in this model (factors 2, 4, and 5) comprise small clusters of energy-related items, while factor 3 mixes four momentum and three energy items, forming a heterogeneous grouping. This structural complexity highlights the interpretive difficulty in multifactor models: while some groupings clearly reflect thematic coherence, others are more ambiguous.

After removing Q16, Q22, and Q23, the four-factor model yields a clearer structure. Two of the factors (factors 2 and 3) form cohesive clusters of energy items, while factor 1 contains a set of momentum items. Factor 3 includes both energy and momentum items, but the overall pattern is more interpretable. This shift suggests that the presence of just a few poorly performing items can disrupt the broader conceptual coherence of the model. Notably, although the factors in the five-factor and four-factor models differ from a statistical perspective, the clustering of items with similar concepts, reflecting momentum and energy differences, emerged as a strong factor in students' reasoning.

To further interpret the dual energy factors, we analyzed the associated items using the conceptual framework of work and mechanical energy proposed by Tong *et al.* [71], which highlights three core concepts: work, conservation of mechanical energy (CME), and the work-energy theorem (WET). Prior research has documented students' difficulties in correctly understanding and relating these concepts. For example, failing to recognize how both force and displacement determine the sign of work [72], struggling to connect changes in potential and kinetic energy [32], or misapplying the work-energy theorem when relating force to changes in kinetic energy [36]. Table III applies the framework proposed by Tong *et al.* to interpret the energy items associated with factors 2 and 3, mainly focusing on the concepts required for the correct answer and the concepts invoked in distractor options. Four of the six items in factor 2 (Q04, Q09, Q15, and Q24) are with distractors unrelated to alternative energy concepts. Q25 also falls into this category, though for a slightly different reason: its correct answer is identified through dimensional analysis, as the first four options have incorrect units. The distractors are likewise not grounded in any alternative energy concepts. The only exception is Q20, which involves multiple concepts. This item presents a two-segment, frictionless slide and asks students to evaluate three statements: one comparing changes in kinetic energy (CME) and two comparing the work done by gravity (work). Although some students may implicitly interpret “work done by gravity” in terms of potential energy, the item nonetheless draws on both CME and work. Taken together, factor 2 is dominated by single-concept items.

By contrast, factor 3 is composed primarily of “multiple-concept” items. Four of its five items (Q06, Q08, Q12, and Q17) include distractors grounded in concepts other than those required for the correct answer. The exception is Q01,

Five-factor Model: Theory-Informed Node Shapes



Four-factor Model: Theory-Informed Node Shapes

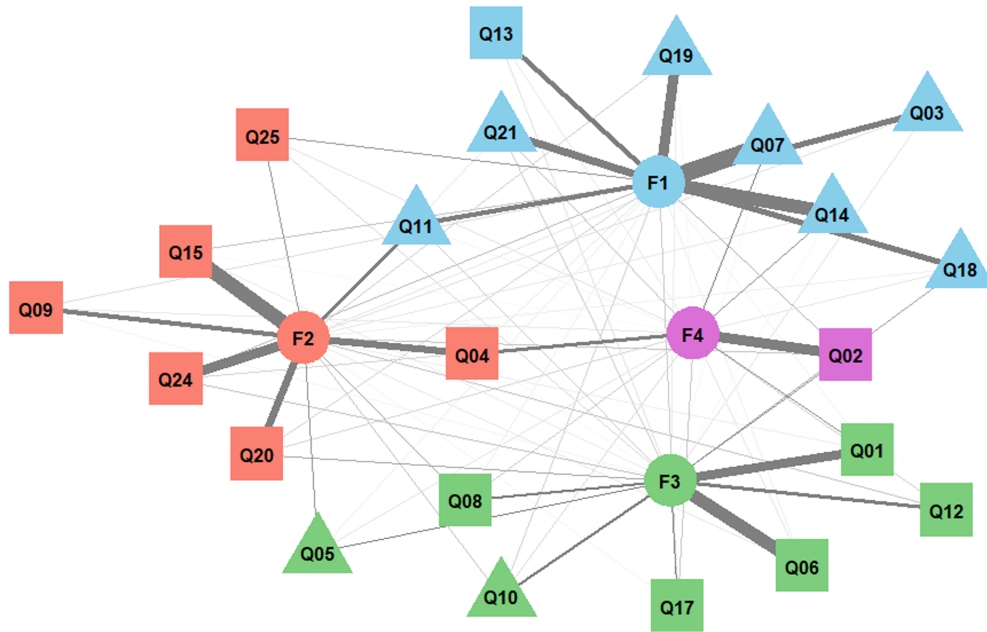


FIG. 2. Network diagrams of the five-factor model [top; five-factor model theory informed node shapes; exploratory factor analysis (EFA) of all 25 Energy and Momentum Conceptual Survey (EMCS) items] and the four-factor model (bottom; four-factor model theory informed node shapes; EFA of the reduced 22 EMCS item set). In the four-factor model, Q16, Q22, and Q23 were removed. In both diagrams, circles denote latent factors (F1–F5 in the top panel; F1–F4 in the bottom panel). Squares represent energy-related items, triangles represent momentum-related ones, classified by the original test design. In the top panel only, a diamond marks item Q16, which was originally classified as relating to both energy and momentum. Node color indicates the factor with the strongest loading for each item, and edge width is proportional to loading magnitude (thicker means stronger). Items are labeled Q01–Q25 (as applicable), and factors are labeled F1–F5 or F1–F4.

TABLE III. Associated energy items for Factors 2 and 3 from the four-factor model, 22 items, showing the concepts required for the correct answer and the concepts underlying distractor options. “Work” refers to the concept of work, “CME” to the concept of conservation of mechanical energy, and “WET” to the concept of the work–energy theorem. For example, “CME (A)” indicates that distractor option A is related to CME. * indicates that the correct answer for the corresponding item is “None of the above.”

Factor	Associated energy items	Correct answer concepts	Distractor concepts
F2	Q04	CME	None
	Q09	CME	None
	Q15	CME	None
	Q20	CME	Work (B, C, D, E)
	Q24	WET	None
	Q25	None*	None
F3	Q01	Work	None
	Q06	Work, WET	CME (A); Work (B, C)
	Q08	CME	Work (D, E)
	Q12	WET	Work (A, B); WET (C)
	Q17	WET	Work (C, D); CME (E)

which asks whether path, speed, and height affect the work done by gravity. From an expert perspective, this item targets only the concept of work. However, students may instead associate “gravitational force” with potential energy or rely on kinematic reasoning prompted by references to speed and height. Overall, the distinction between factors 2 and 3 lies not in performance differences but in whether the items are structured around a single energy concept or require the coordination of multiple concepts. This finding highlights the important role of distractor design: when constructing multiple-choice assessments, whether the distractors represent the same or different concepts can influence the conceptual pathways students activate, even when overall test performance appears unaffected.

C. Number of needed factors

While our best-fitting models revealed five or four factors, it is important to consider the conceptual economy of the EMCS itself. The instrument was designed to assess two fundamental mechanical concepts: energy and momentum. Though one could argue that these concepts fall under a broader mechanics umbrella, collapsing them into a single factor would not yield an informative model. A two-factor model, on the other hand, offers a theoretically grounded benchmark that aligns with the instrument’s design goals.

To examine this possibility, we fit two-factor EFA models to both the 25-item and reduced 22-item datasets. While these models did not achieve the same level of

statistical fit as the five- or four-factor solutions, they showed a strikingly stable conceptual structure (see Fig. 3). In both versions, one factor grouped the core momentum items, along with a few energy items, and the other grouped most of the energy items, with some momentum items joining them. Importantly, removing Q16, Q22, and Q23 had minimal effect on item placement: no items shifted clusters between the full and reduced models. This robustness suggests that the energy-momentum division is conceptually meaningful and statistically stable, even when smaller details in the model fluctuate. Ultimately, while more complex models capture finer distinctions in how students engage with different items, the two-factor model provides a compelling, parsimonious representation of the test’s intended structure. The consistency of these broad groupings across multiple modeling strategies offers support for the EMCS as a conceptually coherent instrument—albeit one that reveals meaningful variation in how students engage with energy concepts in particular.

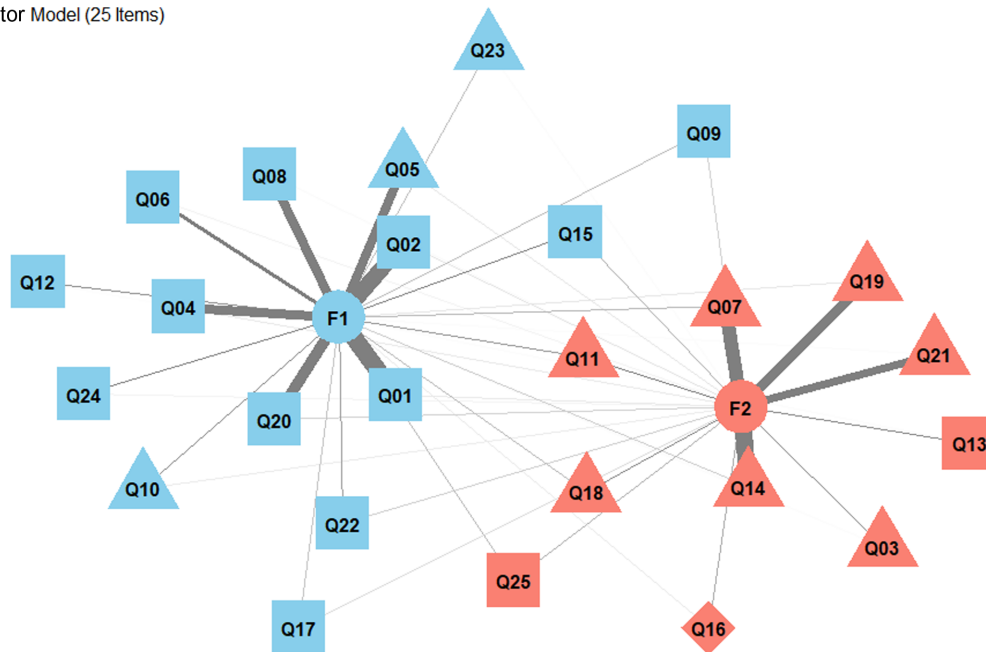
D. Instructional factors behind learning gains

As noted earlier, the post-test averages were relatively low, resulting in small or even negative normalized gains on many items. Here, we provide contextual information about the course and offer some considerations as possible explanations.

The course under investigation was a calculus-based introductory physics sequence taught in a traditional format: two 50-min lectures in a large classroom of over 250 students, one 50-min recitation, and one 110-min laboratory each week. Recitation and lab sections enrolled over 50 students each, with students typically working in groups of three or four. One graduate teaching assistant (GTA) and one undergraduate teaching assistant (UTA) staffed each section. The course was centrally administered, with consistent content and assessments across sections, and weekly coordination meetings involving the course instructor, GTAs, and UTAs. Lectures included occasional clicker questions, but active engagement strategies were not systematically integrated into either lectures or recitations. Exam performance averaged between 70% and 80%, with higher scores on earlier exams covering momentum and energy and lower scores on later exams addressing angular momentum. No other conceptual inventories were administered in this course.

Given this instructional context, the most likely explanation for the negligible learning gains is the limited and inconsistent use of active engagement strategies. The course incorporated elements such as clicker questions, group discussion in recitation, and TA preparation. However, these practices were constrained by the large enrollment and the overall course structure. As a result, the course remained to be rated as “traditional.” Prior work has shown that interactive engagement is associated with substantial improvements on conceptual inventories,

Two-factor Model (25 Items)



Two-factor Model (22 Items)

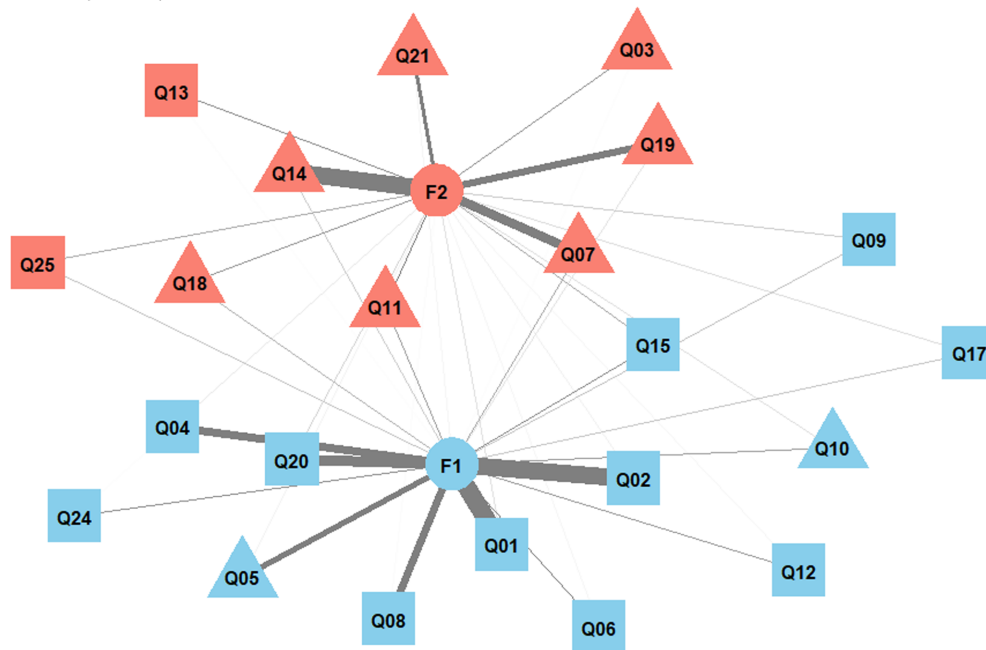


FIG. 3. Network diagrams of the two-factor model based on exploratory factor analysis (EFA) of all 25 (Q01–Q25) Energy and Momentum Conceptual Survey (EMCS) items (above) and the reduced 22-items (below). Shapes, colors, and edge thickness are represented using the same legend as in Fig. 2.

whereas traditional lecture-based approaches typically produce much smaller gains [64].

At the same time, it is important to recognize that “active learning” is not a simple dichotomy. Henderson *et al.* [73,74] have shown that many courses labeled as active-learning environments still devote most of the class time to lecture, suggesting that instructional practice exists on a

spectrum rather than as “active” versus “traditional” categories. Recent work [75,76] has begun to investigate how different active learning methods vary in their effects on conceptual understanding. Building on this perspective, it would be valuable for future studies to include courses typically classified as “traditional” in comparative analyses, in order to more fully capture how a course’s position on

the active-learning spectrum influences student learning gains.

VI. CONCLUSIONS

In this study, we conducted a comprehensive psychometric evaluation of the Energy and Momentum Conceptual Survey (EMCS) by integrating classical test theory (CTT), item response theory (IRT), and exploratory factor analysis (EFA). Rather than relying on any single metric, we evaluated item quality through a holistic lens, considering difficulty, discrimination, and each item's alignment with the overall test structure. Through this multimethod approach, we identified Q16, Q22, and Q23 as consistently weaker in terms of psychometric performance with the current dataset.

Building on this diagnostic phase, we applied EFA using the published guidelines to examine the latent structure of student responses. Removing the three problematic items produced a clearer and more coherent factor model. Notably, two distinct clusters of energy-related items emerged in the revised four-factor model. A close examination of item content revealed that the two energy clusters are primarily distinguished by conceptual scope: one group of items focused on a single energy concept, while the other combined multiple concepts within a single item. This distinction underscores how differences in conceptual demands shape student responses.

On the other hand, we also explored two-factor models, reflecting the EMCS's original design focus on energy and momentum. Although these models did not achieve the same statistical fit as the four- or five-factor alternatives, they consistently yielded a stable conceptual division. This alignment between student performance data and expert-defined categories reinforces the conceptual coherence of the EMCS.

In interpreting the low or negative gains, it is reasonable, on the surface, to classify the course in the current study as “traditional.” Yet the course incorporated substantial active-learning elements, and many courses labeled “active” likewise retain large traditional components. This pattern points to a possible spectrum of active teaching rather than a strict binary. Comparative studies across courses with differing amounts and enactments of active practices would help clarify how a position on this spectrum relates to conceptual learning gains in physics.

V. LIMITATIONS AND DIRECTIONS FOR FUTURE WORK

This study has two primary limitations. First, the dataset lacked student identifiers, which prevented us from matching pre- and post-test data at the individual level. As a result, we could only calculate average item-level normalized gains, limiting our ability to probe why course-level learning gains were small or even negative.

Second, the data were drawn from a single course at a single institution. This narrow context constrained our ability to examine how the instructional environment influences conceptual learning gains and to generalize our findings to broader populations.

To address these limitations, we plan to collect EMCS data from courses at additional institutions. These future datasets will include identifier-linked pre- and post-tests, enabling the calculation of individual-level learning gains and the exploration of how students with different initial understandings progress over time. Moreover, we intend to systematically document instructional contexts, including course syllabi, instructor and TA interviews, and classroom observations, to more precisely situate courses along the active-teaching spectrum. By linking instructional practices with student outcomes across multiple contexts, these follow-up studies will allow us to deepen the psychometric evaluation of the EMCS and advance understanding of how active teaching practices influence conceptual learning in physics.

ACKNOWLEDGMENTS

Our sincere appreciation goes to Samuel Wu, whose cooperation and tolerance during this work were truly remarkable. This work is supported in part by U.S. National Science Foundation Grant No. 2111138.

Opinions expressed are of the authors and not of the Foundation.

DATA AVAILABILITY

The data that support the findings of this article are openly available [77].

APPENDIX

To further illustrate the factor retention decision, results from Horn's parallel analysis are presented in Fig. 4.

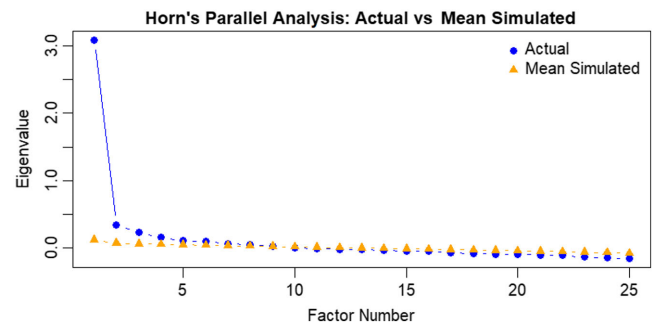


FIG. 4. Horn's parallel analysis results based on principal axis factoring with 100 random permutations. The plot compares the actual eigenvalues from the dataset (blue circles) with the mean eigenvalues from randomly generated datasets of the same size (orange triangles).

This figure provides a visual representation of the point at which the observed eigenvalues fall below the simulated ones, supporting the conclusion that up to nine factors should be retained while acknowledging that parallel analysis offers an upper bound for factor extraction.

-
- [1] S. B. McKagan, L. E. Strubbe, L. J. Barbato, B. A. Mason, A. M. Madsen, and E. C. Sayre, PhysPort use and growth: Supporting physics teaching with research-based resources since 2011, *Phys. Teach.* **58**, 465 (2020).
 - [2] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, *Int. J. Sci. Educ.* **33**, 1289 (2011).
 - [3] T. M. Haladyna and M. C. Rodriguez, *Developing and Validating Test Items* (Routledge, New York, 2013).
 - [4] L. P. McGinness and C. M. Savage, Developing an action concept inventory, *Phys. Rev. Phys. Educ. Res.* **12**, 010133 (2016).
 - [5] J. Li and C. Singh, Developing and validating a conceptual survey to assess introductory physics students' understanding of magnetism, *Eur. J. Phys.* **38**, 025702 (2017).
 - [6] L. Ding and R. Beichner, Approaches to data analysis of multiple-choice questions, *Phys. Rev. ST Phys. Educ. Res.* **5**, 020103 (2009).
 - [7] I. T. Koponen and M. Nousiainen, Concept networks of students' knowledge of relationships between physics concepts: Finding key concepts and their epistemic support, *Appl. Netw. Sci.* **3**, 14 (2018).
 - [8] J. Stewart, C. Zabriskie, S. DeVore, and G. Stewart, Multidimensional item response theory and the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010137 (2018).
 - [9] C. Zabriskie and J. Stewart, Multidimensional item response theory and the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **15**, 020107 (2019).
 - [10] J. Hansen and J. Stewart, Multidimensional item response theory and the brief electricity and magnetism assessment, *Phys. Rev. Phys. Educ. Res.* **17**, 020139 (2021).
 - [11] X. Wu, M. W. Guthrie, E. M. Scanlon, and Y. Li, Analyzing the dimensionality of the energy and momentum conceptual survey using item response theory, *presented at PER Conf. 2003, Sacramento, CA*, 10.1119/perc.2023.pr.Wu.
 - [12] B. Van Dusen and J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, *Phys. Rev. Phys. Educ. Res.* **15**, 020108 (2019).
 - [13] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, *Am. J. Phys.* **84**, 969 (2016).
 - [14] T. Martín-Blas, L. Seidel, and A. Serrano-Fernández, Enhancing Force Concept Inventory diagnostics to identify dominant misconceptions in first-year engineering physics, *Eur. J. Eng. Educ.* **35**, 597 (2010).
 - [15] G. Poutot and B. Blandin, Exploration of students' misconceptions in mechanics using the FCI, *Am. J. Educ. Res.* **3**, 116 (2015).
 - [16] J. Wells, R. Henderson, J. Stewart, G. Stewart, J. Yang, and A. Traxler, Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis, *Phys. Rev. Phys. Educ. Res.* **15**, 020122 (2019).
 - [17] K. L. Malone, Correlations among knowledge structures, Force Concept Inventory, and problem-solving behaviors, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020107 (2008).
 - [18] J. Wang and L. Bao, Analyzing Force Concept Inventory with item response theory, *Am. J. Phys.* **78**, 1064 (2010).
 - [19] M. Planinic, L. Ivanjek, and A. Susac, Rasch model based analysis of the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **6**, 010103 (2010).
 - [20] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, *Am. J. Phys.* **80**, 825 (2012).
 - [21] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020105 (2012).
 - [22] P. Eaton and S. D. Willoughby, Confirmatory factor analysis applied to the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010124 (2018).
 - [23] J. Docktor, K. Heller, C. Henderson, M. Sabella, and L. Hsu, Gender differences in both Force Concept Inventory and introductory physics performance, *AIP Conf. Proc.* **1064**, 15 (2008).
 - [24] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **14**, 010103 (2018).
 - [25] J. B. Buncher, J. M. Nissen, B. Van Dusen, and R. M. Talbot, Is the Force Concept Inventory biased across the intersections of gender and race?, *Phys. Rev. Phys. Educ. Res.* **21**, 010137 (2025).
 - [26] P. Nieminen, A. Savinainen, and J. Viiri, Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, *Phys. Rev. ST Phys. Educ. Res.* **6**, 020109 (2010).
 - [27] P. Nieminen, A. Savinainen, and J. Viiri, Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning, *Phys. Rev. ST Phys. Educ. Res.* **8**, 010123 (2012).
 - [28] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998).

- [29] T. I. Smith and M. C. Wittmann, Applying a resources framework to analysis of the force and motion conceptual evaluation, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020101 (2008).
- [30] R. K. Thornton, D. Kuhl, K. Cummings, and J. Marx, Comparing the force and motion conceptual evaluation and the Force Concept Inventory, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010105 (2009).
- [31] J. Wells, R. Henderson, A. Traxler, P. Miller, and J. Stewart, Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis, *Phys. Rev. Phys. Educ. Res.* **16**, 010121 (2020).
- [32] C. Singh and D. Rosengrant, Multiple-choice test of energy and momentum concepts, *Am. J. Phys.* **71**, 607 (2003).
- [33] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics*, 1st ed. (Prentice Hall, Upper Saddle River, NJ, 2002).
- [34] R. W. Chabay and B. A. Sherwood, *Matter & Interactions*, 4th ed. (Wiley, Hoboken, NJ, 2015).
- [35] <https://www.physport.org/>.
- [36] R. A. Lawson and L. C. McDermott, Student understanding of the work-energy and impulse-momentum theorems, *Am. J. Phys.* **55**, 811 (1987).
- [37] B. A. Lindsey, P. R. L. Heron, and P. S. Shaffer, Student ability to apply the concepts of work and energy to extended systems, *Am. J. Phys.* **77**, 999 (2009).
- [38] N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello, and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, *J. Eng. Educ.* **104**, 454 (2015).
- [39] L. Plonsky, *Advancing Quantitative Methods in Second Language Research*, 1st ed. (Taylor & Francis Group, Oxford, 2015).
- [40] M. C. Howard, A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?, *Int. J. Hum. Comput. Interact.* **32**, 51 (2016).
- [41] M. W. Watkins, Exploratory factor analysis: A guide to best practice, *J. Black Psychol.* **44**, 219 (2018).
- [42] M. C. Howard, A systematic literature review of exploratory factor analyses in management, *J. Bus. Res.* **164**, 113969 (2023).
- [43] M. J. Brundage, A. Maries, and C. Singh, Using the Energy and Momentum Conceptual Survey to investigate progression in student understanding from introductory to advanced levels, *Phys. Rev. Phys. Educ. Res.* **19**, 020132 (2023).
- [44] R. F. DeVellis, Classical test theory, *Med. Care* **44**, S50 (2006).
- [45] P. Eaton, K. Johnson, B. Frank, and S. Willoughby, Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism, *Phys. Rev. Phys. Educ. Res.* **15**, 010102 (2019).
- [46] E. E. Cureton, The upper and lower twenty-seven per cent rule, *Psychometrika* **22**, 293 (1957).
- [47] R. L. Doran, Implications for measurement and evaluation from the trends of science education, *Sci. Educ.* **60**, 199 (1976).
- [48] T. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation* (SAGE Publications, Inc., Thousand Oaks, CA, 2005).
- [49] M. Tavakol and R. Dennick, Making sense of Cronbach's alpha, *Int. J. Med. Educ.* **2**, 53 (2011).
- [50] *Handbook of Modern Item Response Theory*, edited by W. J. Van Der Linden and R. K. Hambleton (Springer, New York, NY, 1997).
- [51] J. P. Stevens, *Applied Multivariate Statistics for the Social Sciences*, 5th ed. (Taylor and Francis, London, 2009).
- [52] M. S. Bartlett, Tests of significance in factor analysis, *Br. J. Stat. Psychol.* **3**, 77 (1950).
- [53] H. F. Kaiser, A second generation little jiffy, *Psychometrika* **35**, 401 (1970).
- [54] C. D. Dziuban and E. C. Shirkey, When is a correlation matrix appropriate for factor analysis? Some decision rules., *Psychol. Bull.* **81**, 358 (1974).
- [55] R. K. Henson and J. K. Roberts, Use of exploratory factor analysis in published research: Common errors and some comment on improved practice, *Educ. Psychol. Meas.* **66**, 393 (2006).
- [56] A. B. Costello and J. Osborne, Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis, *Pract. Assess. Res. Eval.* **10**, 7 (2005).
- [57] J. M. Conway and A. I. Huffcutt, A review and evaluation of exploratory factor analysis practices in organizational research, *Organ. Res. Meth.* **6**, 147 (2003).
- [58] R Core Team, *r: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2024).
- [59] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, and Posit Software, PBC, *DPLYR: A grammar of data manipulation* (2023), <https://dplyr.tidyverse.org/>.
- [60] H. Wickham, L. Henry, and Posit Software, PBC, *PURRR: Functional programming tools* (2025), <https://purrr.tidyverse.org/>.
- [61] W. Revelle, *PSYCH: Procedures for psychological, psychometric, and personality research* (2025), <https://cran.r-project.org/package=psych>.
- [62] P. Chalmers *et al.*, MIRT: A multidimensional item response theory package for the R environment, *J. Stat. Softw.*, **48**, 1 (2012).
- [63] S. Epskamp, G. Costantini, J. Haslbeck, A. Isvoranu, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom, *QGRAPH: Graph plotting methods, psychometric data visualization and graphical model estimation* (2023), <https://cran.r-project.org/package=qgraph>.
- [64] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
- [65] J. D. Marx and K. Cummings, Normalized change, *Am. J. Phys.* **75**, 87 (2007).
- [66] L. Bao, Theoretical comparisons of average normalized gain calculations, *Am. J. Phys.* **74**, 917 (2006).
- [67] J. M. Nissen, R. M. Talbot, A. Nasim Thompson, and B. Van Dusen, Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories, *Phys. Rev. Phys. Educ. Res.* **14**, 010115 (2018).

- [68] H. F. Kaiser, An index of factorial simplicity, *Psychometrika* **39**, 31 (1974).
 - [69] R. Driver and L. Warrington, Students' use of the principle of energy conservation in problem situations, *Phys. Educ.* **20**, 171 (1985).
 - [70] J. Tuminaro and E. F. Redish, Elements of a cognitive model of physics problem solving: Epistemic games, *Phys. Rev. ST Phys. Educ. Res.* **3**, 020101 (2007).
 - [71] D. Tong, J. Liu, Y. Sun, Q. Liu, X. Zhang, S. Pan, and L. Bao, Assessment of student knowledge integration in learning work and mechanical energy, *Phys. Rev. Phys. Educ. Res.* **19**, 010127 (2023).
 - [72] R. Driver and L. Warrington, Students' use of the principle of energy conservation in problem situations, *Phys. Educ.* **20**, 171 (1985).
 - [73] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020104 (2012).
 - [74] M. Dancy, C. Henderson, N. Apkarian, E. Johnson, M. Stains, J. R. Raker, and A. Lau, Physics instructors' knowledge and use of active learning has increased over the last decade but most still lecture too much, *Phys. Rev. Phys. Educ. Res.* **20**, 010119 (2024).
 - [75] M. Sundstrom, J. Gambrell, C. Green, A. L. Traxle, and E. Brewe, Relative benefits of different active learning methods to conceptual physics learning, [arXiv:2505.04577](https://arxiv.org/abs/2505.04577).
 - [76] I. Bukola, M. Sundstrom, J. Gambrell, O. Ross, A. L. Traxler, and E. Brewe, Multi-institutional assessment of peer instruction implementation and impacts using the framework for interactive learning in lectures, [arXiv:2508.08422](https://arxiv.org/abs/2508.08422).
 - [77] X. Wu, Master Data File (2025), [10.17605/OSF.IO/KV9PA](https://doi.org/10.17605/OSF.IO/KV9PA).
- Correction:* The last author's name was presented incorrectly and has been fixed.