# Approaches to data analysis of multiple-choice questions

Lin Ding[1] and Robert Beichner[2]
[1]*Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*
[2]*Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA*
(Received 12 February 2009; published 10 September 2009)

This paper introduces five commonly used approaches to analyzing multiple-choice test data. They are classical test theory, factor analysis, cluster analysis, item response theory, and model analysis. Brief descriptions of the goals and algorithms of these approaches are provided, together with examples illustrating their applications in physics education research. We minimize mathematics, instead placing emphasis on data interpretation using these approaches.

DOI: 10.1103/PhysRevSTPER.5.020103 PACS number(s): 01.40.Fk, 01.40.gf, 01.40.G−

## I. INTRODUCTION

Multiple-choice tests are increasingly used in physics education to assess student learning. Appropriate and effective approaches to data analysis of multiple-choice tests thus become an important research topic. To facilitate data analysis and interpretation, physics education researchers have adopted various testing techniques from educational and psychological studies. These techniques benefited many studies published in this journal and others such as the *American Journal of Physics*.

Despite voluminous literature on mathematical theories of diverse testing techniques, a concise introduction to frequently encountered approaches of data analysis suitable for physics education research (PER) is much needed. In this paper, we briefly introduce five approaches to analyzing multiple-choice test data; these are classical test theory, factor analysis, cluster analysis, item response theory, and model analysis (see Table I). Specifically, we introduce the goals and basic algorithms of each approach, offering examples to demonstrate how each approach can be used for data interpretation. Emphasis is placed on applications of these approaches in the context of PER studies. Since it is *not* our intention to present comprehensive theories of statistics, we minimize mathematical details and avoid derivations that can be found in the listed references. We also do *not* intend to pursue highly technical issues that are controversial even among statisticians and psychometricians; therefore notions, terminologies, names, and discussions presented in this paper are in compliance with conventional norms that are commonly recognized.

Other related issues not covered herein include the pros and cons of multiple-choice tests,[1] various types of test "validity,"[2] and pre/post use of multiple-choice tests to gauge the effectiveness of traditional and reform courses.[3]

The remainder of the paper is organized into six sections, five of which are devoted to the five approaches, respectively (Secs. II–VI), followed by a brief summary (Sec. VII).

## II. CLASSICAL TEST THEORY

Classical test theory is an important part of the foundation of modern measurement theory.[4] It assumes that a total test score is made up of two components: true score and random

TABLE I. Five approaches to analyzing multiple-choice test data.

| | | Goal or purpose | Basic algorithm |
|---|---|---|---|
| Classical test theory | | Evaluate item or test reliability and discriminatory power | Perform item analysis and test analysis |
| Factor analysis | Principal component analysis | Reduce the number of variables | Solve eigenvalue equations for correlation matrix |
| | Common factor analysis | Explore underlying factors | Solve eigenvalue equations for adjusted correlation matrix |
| Cluster analysis | | Classify subjects into groups | Calculate Euclidian distances and merge/divide subjects |
| Item response theory | | Estimate item characteristics and subjects' latent abilities | Use logistic functions to formulate data |
| Model analysis | | Represent probabilities of using different models | Calculate density matrix and solve eigenvalue equations |

 

error. Based on this assumption, classical test theory gives rise to a number of statistical analyses for test evaluation, including item analysis and test analysis.[5] The purpose of these analyses is to examine if a test is reliable and discriminating. For a reliable test, similar outcomes are expected if the test is administered twice (at different times), assuming the examinees' performance is stable and the testing conditions are the same. For a discriminating test, results can be used to clearly distinguish those who have a robust knowledge of tested materials from those who do not. In the following, we provide brief introductions to both item analysis and test analysis. One can follow these analyses to perform test evaluations.

Item analysis encompasses three measures: item difficulty level ($P$), discrimination index ($D$), and point biserial coefficient ($r_{\mathrm{pbi}}$). Item difficulty is a measure of the easiness of an item (although it is ironically called "item difficulty" level) and is defined as the proportion of correct responses,

$$P = N_1/N.$$

Here $N_1$ is the number of correct responses and $N$ is the total number of students taking the test. Ideally, items with difficulty level around 0.5 have the highest reliability because questions that are either extremely difficult or extremely easy do not discriminate between students (since nearly everyone gets them wrong or right). Practically, item difficulty values ranging from 0.3 to 0.9 are acceptable.[6] If items with extremely low or extremely high difficulty values are detected, one may consider revising these items to make them easier (or more difficult).

The item discrimination index measures how powerful an item is in distinguishing high-achieving students from low-achieving students. It is defined as the difference in percentages of correct response to an item between the top quartile and the bottom quartile students,[7]

$$D = (N_{\mathrm{H}} - N_{\mathrm{L}})/(N/4).$$

Here $N_{\mathrm{H}}$ and $N_{\mathrm{L}}$ are the numbers of correct responses in the top quartile and bottom quartile, respectively, and $N$ is the total number of students. "Quartile" can be determined by using either an "internal" criterion (students' scores on the test being considered) or an "external" criterion (e.g., students' grade point averages).[6] The criterion used in most PER discussions is internal. Occasionally researchers use top and bottom thirds or some other division of scores, depending on what best suits their needs. If an item is discriminative, one can expect the number of correct responses in the top quartile ($N_{\mathrm{H}}$) to be much greater than that ($N_{\mathrm{L}}$) in the bottom quartile, thus a high discrimination index. A commonly adopted standard[8] for a satisfactory item discriminatory index is $D \geq 0.3$. Higher values are better. In case of a low item discriminatory index, one may need to scrutinize the item to see if the statement of the question is clear. Sometimes, a poorly worded item can cause strong students to overthink the question, posing a negative effect on their performance and thus lowering the item discrimination index. Another possible situation for low item discrimination is when an item is either too difficult (low difficulty index) or too easy (high difficulty index). In these cases, the difference in performance between the top quartile and bottom quartile students is small; hence the item discrimination index is low.

The point biserial coefficient is a measure of individual item reliability and is defined as the correlation between the item scores and total scores,[9]

$$r_{\mathrm{pbi}} = \frac{\overline{X_1} - \overline{X_0}}{\sigma_x}\sqrt{P(1-P)}.$$

Here, $\overline{X_1}$ is the average total score for those who correctly answer an item, $\overline{X_0}$ is the average total score for those who incorrectly answer the item, $\sigma_x$ is the standard deviation of total scores, and $P$ is the difficulty index for this item. A reliable item should be consistent with the rest of the test, so a fairly high correlation between the item score and the total score is expected. A satisfactory point biserial coefficient[10] is $r_{\mathrm{pbi}} \geq 0.2$. Once again, higher values are better. If an item shows a low biserial coefficient, it indicates this item may not test the same material (or may not test the material at the same level) as other items. Revisions may be considered to make this item more comparable to the rest of the test.

Test analysis has two measures: Kuder-Richardson reliability index ($r_{\mathrm{test}}$) and Ferguson's delta ($\delta$). These two measures are used to evaluate an entire test (rather than evaluate individual items). Kuder-Richardson reliability measures the internal consistency of a test. In other words, it examines whether or not a test is constructed of parallel items that address the same materials. Higher correlations among individual items result in a greater Kuder-Richardson index, indicating higher reliability of the entire test. For a multiple-choice test where each item is scored as "correct" or "wrong," the reliability index is calculated as follows:[11,12]

$$r_{\mathrm{test}} = \frac{K}{K-1}\left(1 - \frac{\Sigma\, P_i(1-P_i)}{\sigma_x^2}\right).$$

This formula is known as KR-20 after the equation number in the famous Kuder and Richardson paper.[11] Here, $K$ is the number of the test items, $P_i$ is the difficulty level for the $i$th item, and $\sigma_x$ is the standard deviation of total score. A widely accepted criterion is that a test of reliability higher than 0.7 is considered reliable for group measurement. An $r_{\mathrm{test}}$ value greater than 0.8 is the rule of thumb indicating a test is suitable for use in assessing individuals.[13] If a low reliability index is detected, one may first consider examining items with low discrimination index and low point biserial coefficient. Because these items often are not consistent with other items, they can negatively impact the reliability of the entire test.

Ferguson's delta measures the discriminatory power of an entire test. Specifically, it investigates how broadly students' total scores are distributed over the possible range. Generally, the broader the score distribution is, the better the test is in discriminating among students at different levels. The calculation of Ferguson's delta is given as follows:[14,15]

$$\delta = \frac{N^2 - \Sigma\, f_i^2}{N^2 - N^2/(K+1)}.$$

Here, $N$ is the total number of students taking the test, $K$ is the number of test items, and $f_i$ is the number of students

TABLE II. Evaluations of the MIET.

| Test statistics | MIET values | Desired values |
|---|---|---|
| Difficulty index | Average of 0.49 | [0.30, 0.90] |
| Discrimination index | Average of 0.38 | ≥0.30 |
| Point biserial coefficient | Average of 0.33 | ≥0.20 |
| Reliability index | 0.74 | ≥0.70 |
| Ferguson's delta | 0.97 | ≥0.90 |

whose total score is $i$. Generally, if a test has Ferguson's delta greater than 0.90, it is considered to provide good discrimination among students.[16]

The above analyses are easy to perform. Once one obtains basic descriptive statistics such as total scores and item scores, calculating these indices or coefficients becomes fairly straightforward. With these analysis results at hand, one then can evaluate a test regarding its reliability and discrimination. For example, Table II shows the analysis results of a 33-item research-based energy assessment designed for the Matter & Interactions mechanics course, namely, the Matter & Interactions Energy Test (MIET).[17] As seen, MIET is a medium difficult test, and it has satisfactory overall discrimination and reliability for both the individual items and the entire test. Interested readers can replicate these results from detailed data in Appendix A by following the aforementioned formulations. More studies using these analyses for physics assessment evaluation can be found elsewhere, for example, evaluation of the Test of Understanding Graphs in Kinematics (TUG-K),[18] a Conceptual Survey of Electricity and Magnetism (CSEM),[19] a Brief Electricity and Magnetism Assessment (BEMA),[15] a multiple-choice test of energy and momentum concepts,[20] and the Determining and Interpreting Resistive Electric Circuits Test (DIRECT).[21]

## III. FACTOR ANALYSIS

Factor analysis is a short name for factor analytical techniques, and it includes both *principal component analysis* (PCA) and *common factor analysis*.[22] Generally, factor analysis is performed when one has a large number of observed variables but wishes to reduce it to conveniently explain data. For example, one administers a 30-item multiple-choice physics test among several hundred students; he or she then collects a data set of 30 variables (each variable corresponding to an individual item). Going through all 30 variables in detail is time consuming, and the abundance of data may not clearly reveal the overall picture of the results. A small number of artificial variables that account for the most (co)variance in the observed variables can facilitate data interpretation. Factor analysis is a statistical technique that constructs a small set of artificial variables through linear combinations of highly correlated observed variables. These artificial variables are called "components" in principal component analysis or "common factors" in common factor analysis. Although principal component analysis and common factor analysis share similar features, they serve fundamentally different purposes.[23] The former aims to develop components without assuming a causal relationship between components and observed variables, whereas the latter explores underlying "factors" which are assumed to impose causal effects on observed variables. Details can be found in the following sections.

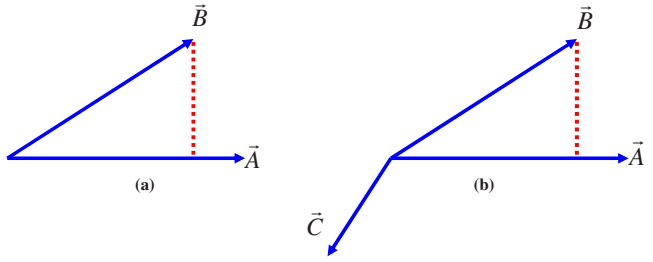


FIG. 1. (Color) Vector representation of observed variables: (a) correlation of two vectors can be represented as the projection of one vector onto the other; (b) three dimensions are needed to represent three vectors.

Here we start with a geometrical view of factor analysis to help readers visualize this technique. Then we discuss some basics of PCA and common factor analysis with examples demonstrating how each approach can be used in multiple-choice test data analysis. Some practical issues are addressed, and the underlying differences between PCA and common factor analysis are emphasized.

### A. Geometrical view

Nichols[24] described a helpful geometrical analogy to factor analysis. Imagine each observed variable being represented by a vector of unit length. The correlation between any two variables is represented by the cosine of the angle between the two vectors. (Since we use unit vectors, this is equivalent to the projection of one vector onto the other [see Fig. 1(a)]). With two nonparallel vectors, we need two dimensions to represent them. If we have a third vector jutting off the paper, we will need a third dimension [see Fig. 1(b)]. This idea can be extended to any number of vectors in a multidimensional space. To simplify our analogy without losing generality, we consider three dimensions by assuming there are only three factors. The task of factor analysis is to find three dimensions that best summarize observed variables represented by blue vectors in Fig. 2.

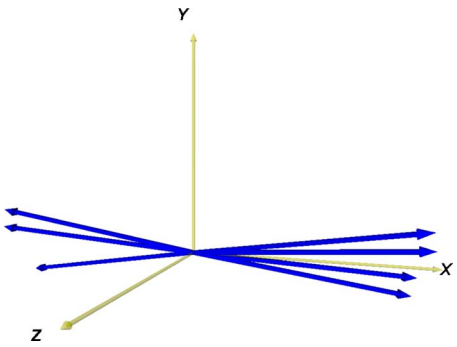


FIG. 2. (Color) A starburst configuration of observed variables is represented by blue arrows.

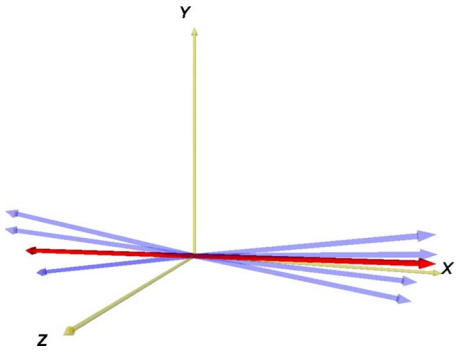


FIG. 3. (Color) Sum of the squared projections of the blue arrow onto the red arrow (through the origin) is maximized. The red arrow represents factor 1.

To find the first factor, we add a bidirectional vector through the origin and spin it around until we have maximized the sum of the squared projections of the blue vectors onto it. The red arrow in Fig. 3 represents the first factor. The projections of the variables (blue vectors) onto the factor are referred to as the factor loadings of factor 1.

Now place a flat screen at the origin, perpendicular to the first factor (red arrow). Shine a light toward the origin from beyond each end of the red vector and cast shadows of the blue vectors onto the screen (see Fig. 4). These shadows are the residual correlations with the first factor partialled out.

We now place a new bidirectional green vector on the screen through the origin and perpendicular to the red vector. We then rotate the green vector until the sum of the squared projections of the shadows onto it (not the blue vectors, but their shadows on the screen) is maximal. The green bidirectional vector in Fig. 5 represents the second factor. Since the green vector is perpendicular to the first factor, factor 1 and factor 2 are orthogonal.

We now place a third bidirectional orange vector on the screen through the origin and perpendicular to the red and green vectors (see Fig. 5). The projections of the shadows on the orange vector are called the loadings on the third factor.

By finally clearing away all the extraneous representations, we see how the three factor vectors form a coordinate system that is more closely aligned with the original vectors than the $x$, $y$, and $z$ axes. Figure 6 shows the view looking down the factor 1 axis toward the origin. Although impossible to visualize with more than three factors, the mathematical approach and meaning are the same regardless of how many dimensions are involved.

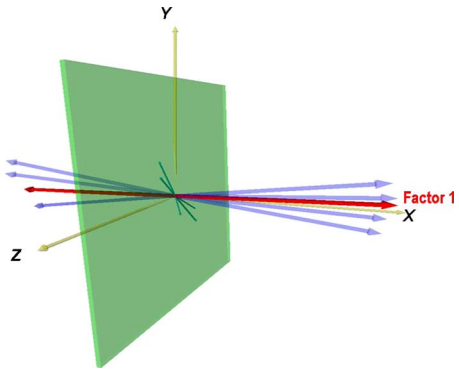


FIG. 4. (Color) Cast projections of the blue arrows onto the green screen perpendicular to factor 1.

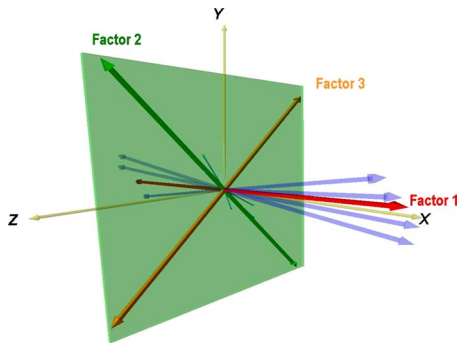


FIG. 5. (Color) The sum of the squared projections of the green shadows onto the green vector is maximized. The green vector presents factor 2. The orange vector is perpendicular to both the red and green vectors, representing factor 3. In a general case of multiple dimensions, a rotation of all factors probably is desirable and if needed can still maintain orthogonality among them.

### B. Principal component analysis

PCA is a variable reduction procedure. Suppose one obtains a data set of 30 variables from administering a 30-item multiple-choice test. Usually several items test the same topic, and thus students' scores on these items should ideally display high correlations. PCA groups these highly correlated items together and collapses them into a small number of components through weighted linear combinations as shown below:

$$C_1 = b_{11}Q_1 + b_{12}Q_2 + b_{13}Q_3 + \cdots + b_{1,30}Q_{30},$$

$$C_2 = b_{21}Q_1 + b_{22}Q_2 + b_{23}Q_3 + \cdots + b_{2,30}Q_{30},$$

$$\cdots.$$

Here $C$'s are components, $Q_i$ represents the $i$th item in the test, and $b$'s are weights of individual items. The main task of PCA is to optimize the weights ($b$ values) so that these components constructed as such account for variance in the

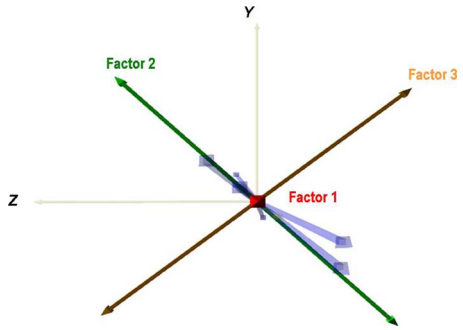


FIG. 6. (Color) Three factors are more aligned with the blue vectors than the $x$, $y$, and $z$ axes. Here, the $x$ axis is too close to factor 1 and therefore is not labeled in the picture.

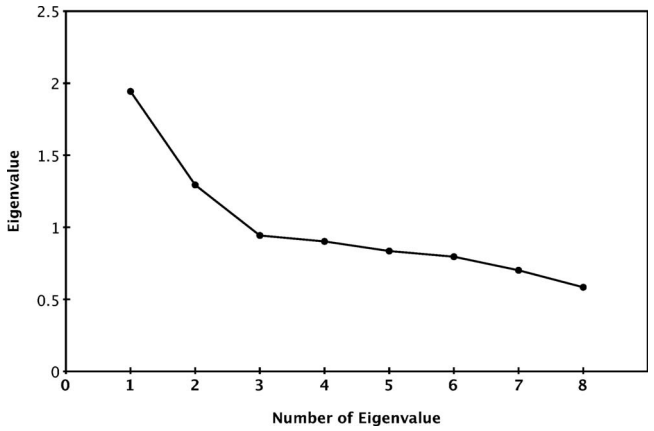


FIG. 7. A "scree plot" for PCA of eight MIET items.

observed data more than components constructed otherwise. PCA solves eigenvalue equations for the optimal weights from a correlation matrix.

The following is an example, showing PCA results on eight selected questions (Q12–Q14, Q16, Q27, and Q31–Q33) from MIET.[17] A correlation matrix for these eight questions is calculated using 389 students' responses and is provided in Appendix A for interested readers to replicate the results. From the $8\times8$ matrix, PCA first solves for eight eigenvalues and corresponding eigenvectors (components). Figure 7 shows a "scree plot"[25] of these eight eigenvalues. (The name of the plot refers to the rubble or scree that collects at the bottom of a cliff.)

As seen, the curve drops quickly, and at the third eigenvalue its rate of decrease reduces significantly. Because the eigenvalues indicate the amounts of variance explained by each eigenvector (component), it is conceivable that the first two components may be enough to explain a meaningful amount of variance in the observed data. Thus, the first two components are retained. The components obtained heretofore can be further transformed into either correlated or uncorrelated components through orthogonal or oblique rotations[26,27] at the researcher's discretion. (Orthogonal rotation rotates all components as a rigid body system, preserving the orthogonality among them, whereas oblique rotation does not maintain the orthogonality.) Since either rotation may generate more sensible components than the other, no rigid rule exists as for which one is more preferable. Practically, one may consider trying both rotations to explore which facilitates data interpretation better. Interested readers can see Refs. 26,27 for details. For simplicity, we consider two uncorrelated (orthogonal) components and use SAS to obtain component loadings in Table III. Here, components are the eigenvectors of a correlation matrix. Component loadings are the linear regression coefficients of the components for each variable, which are calculated from the aforementioned linear equations in this section.

It is clear from Table III that the first four questions have large loadings on component 1 and small loadings on component 2. Conversely, the last four questions show a reverse pattern. Thus it is reasonable to conclude that the first four questions can be summarized by component 1 and the last four questions by component 2. It is now the researcher's job to describe what each component means. In this example, a close inspection of these questions reveals that the first questions test basic concepts of energy in bound or unbound systems and the last four questions test the work-energy theorem (see Appendix B for the eight selected questions). Therefore, it is sensible to name the first component "energy in bound or unbound system" and the second component "the work-energy theorem."

TABLE III. Component loadings for PCA of eight MIET items. An asterisk indicates an "important" factor loading as suggested by Hair *et al.* (Ref. 28).

| | Component 1 | Component 2 | Question content |
|---|---|---|---|
| Q12 | 0.66* | 0.08 | Energy graph, bound or unbound system |
| Q13 | 0.73* | 0.03 | Energy graph, bound or unbound system |
| Q14 | 0.63* | −0.12 | Energy graph, bound or unbound system |
| Q16 | 0.54* | 0.02 | Energy of a bound system |
| Q27 | 0.09 | 0.53* | Work-energy theorem |
| Q31 | −0.11 | 0.63* | Work-energy theorem |
| Q32 | 0.12 | 0.59* | Work-energy theorem |
| Q33 | −0.06 | 0.70* | Work-energy theorem |

The simple pattern of component loadings in the preceding example is desirable for data interpretation. When judging the size of component loadings, one may use Hair's suggestion[28] to consider loadings of $\pm0.3$ as minimal, $\pm0.4$ as more important, and $\pm0.5$ as significant. That said, there really is no rigid criterion to follow, and it often relies on the researcher's experience to decide what cutoff value to use. Some software such as SAS[29] has built-in functions to flag significant loadings on a case-by-case basis instead of using a "one-size-fit-all" criterion.

### C. Common factor analysis

Common factor analysis assumes some unmeasured "factors" as the underlying causes of observed variables.[30,31] Suppose one administers four multiple-choice questions among students and collects a data set of four variables. With an assumption that two common factors underlie the observed data, the relations among these two common factors and four questions can be depicted in Fig. 8. Here, rectangu-

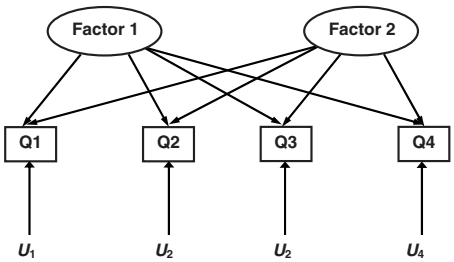


FIG. 8. Common factors and unique factors in common factor analysis.

lar boxes represent observed variables (questions 1–4) and elliptical circles indicate latent (unmeasured) common factors. Single-headed arrows signify causal influences on observed variables. As shown, each observed variable receives influences from not only common factors (factors 1 and 2) but also its own underlying factor ($U$). As opposed to the common factors, these $U$ factors are called unique factors. The goal of common factor analysis is to identify the common factors to account for the covariance among observed variables.

As seen in Fig. 8, each observed variable can be expressed as follows:

$$Q_1 = b_{11}F_1 + b_{12}F_2 + U_1,$$

$$Q_2 = b_{21}F_1 + b_{22}F_2 + U_2.$$

Here $Q$'s are observed variables, $F$'s are common factors, $U$'s are unique factors, and $b$'s are regression coefficients (weights). Similarly to PCA, common factor analysis calculates optimal weights ($b$ values) by solving eigenvalue equations of the observed correlation matrix. However, the diagonal elements of the correlation matrix are no longer 1's. Rather, they are replaced with variances that are accounted for by common factors. In other words, the portion of variances explained by unique factors is discarded. This is a major difference between PCA and common factor analysis. Here, this new correlation matrix is called adjusted correlation matrix.[32] Since the portion of variances due to common factors cannot be known *a priori*, it is practical to use estimates for an initial trial. Often commercial software has functions to generate estimates, followed by iterations to solve the equations.

In the following we provide an example of common factor analysis using the same data as for PCA. The purpose of this example is to uncover the underlying structures of measured variables. (Here we do not predetermine possible factors. Such approach is known as exploratory factor analysis. In cases where common factor analysis is used to confirm predetermined factors, it is called confirmatory factor analysis. Readers can refer to Ref. 22 for confirmatory factor analysis.) The initial adjusted correlation matrix is included in Appendix A for interested readers to replicate the results. We choose orthogonal rotation and the maximum likelihood method[29] (an iteration method in which parameter estimates are those most likely to have yielded the observed correlation matrix) in SAS for common factor analysis. Other iteration methods include unweighted least squares (minimizing the sum of squared differences between estimated and observed correlation matrices) and generalized least squares (adjusting the unweighted least squares by weighing the correlations inversely to their unique variables, $U$'s).[29] Although maximum likelihood may be the most frequently used method, there are no dogmatic criteria as for which one is more preferable than others. A rule of thumb is to choose a method that best facilitates data interpretation. The reason we choose maximum likelihood is that for our data this method produces more sensible and easy-to-interpret results. In Table IV, the final factor loadings for each question are shown. As previously stated, SAS has built-in functions to flag significant loadings. Loadings with an asterisk in Table IV are considered significant by SAS.

TABLE IV. Factor loadings for common factor analysis of eight MIET items.

| | Factor 1 | Factor 2 | Content |
|---|---|---|---|
| Q12 | 0.54* | 0.13 | Energy graph, bound or unbound system |
| Q13 | 0.66* | 0.08 | Energy graph, bound or unbound system |
| Q14 | 0.35* | 0.12 | Energy graph, bound or unbound system |
| Q16 | 0.31 | 0.02 | Energy of a bound system |
| Q27 | 0.14 | 0.33* | Work-energy theorem |
| Q31 | 0.14 | 0.35* | Work-energy theorem |
| Q32 | 0.15 | 0.44* | Work-energy theorem |
| Q33 | 0.04 | 0.53* | Work-energy theorem |

As seen, the results are very similar to those obtained from PCA. However, factor loadings in Table IV are generally lower than component loadings in PCA (Table III). This is understandable because common factor analysis considers only the portion of variances explained by common factors, whereas PCA analyzes the total variances in observed data without assuming underlying effects. This is why the purpose of PCA is purely data reduction, while the purpose of common factor analysis is to understand causes.[33,34]

Another interesting difference between the above results and those in Table III is manifested in Q16. Previously, PCA generates a high loading on component 1 "energy in bound or unbound system." Here the loading on factor 1 is barely above the minimum suggested by Hair *et al.*[28] In fact, SAS does not even flag this loading as significant. A closer inspection reveals that Q16 differs from the first three questions in that it is not formatted in graphical representations. This difference may have lowered its loading value on factor 1. Put differently, there may exist factors other than the above two that can better explain the observed data for question 16.

### D. Practical issues

Both of the above two examples yield a simple pattern of loadings for observed variables; that is, nearly all the questions have loadings high on one component (or factor) and low on others. This simple pattern makes it easy to interpret the meaning of components or factors. In reality, one rarely obtains such a simple pattern, especially when dealing with binary data (1 or 0) collected from dichotomously scored multiple-choice questions. In that case, one can start with correlations for dichotomous[35] variables, as suggested by Heller and Huffman,[36] instead of Pearson correlations. Another approach is "reduced basis factor analysis"[37] proposed by Adams *et al.* In this approach, factor analysis is repeatedly performed by adjusting the initial factors so as to maximally comprise between predetermined results and the results acquired from real data.

Results obtained from either PCA or common factor analysis should be used heuristically and not in an absolute manner. Oftentimes one may find PCA generates results more interpretable than common factor analysis and vice

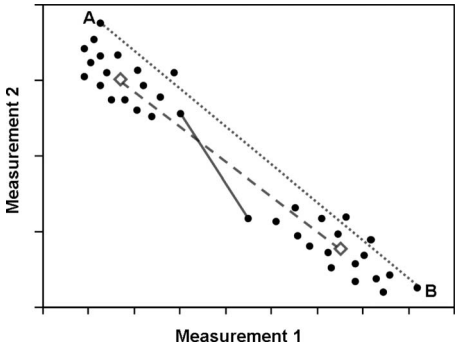


FIG. 9. Euclidian distance as a measure of similarity. The dotted line connects the farthest apart points of two clusters (complete linkage). The solid line connects the closest points of two clusters (single linkage). The dashed line connects the centroids (diamonds) of two clusters. The average linkage is not shown, as it is a mathematical average of all pairwise distances between two clusters.

versa, or orthogonal rotation yields loading patterns simpler than oblique rotation and vice versa. Researchers need to make decisions based on practical needs for the best interpretation of data. Analysis of factor analysis results is sometimes more art than science.

## IV. CLUSTER ANALYSIS

Cluster analysis[38] is used when one wants to see if students can be classified into different groups with distinctive characteristics. For example, one administers a 30-item multiple-choice test to 300 students. In addition to just looking over students' total scores, you may also want to know if students display any distinctive response patterns. In this case, cluster analysis is an appropriate tool to examine the similarities (or dissimilarities) among students' responses and thus to group them into different clusters.

Cluster analysis often uses Euclidian distances to measure similarities between any two subjects.[39,40] [Here, we confine our discussion to interval (continuous) data. For binary data (1 or 0), one may consider using other measures, such as binary Lance and Williams nonmetric distances.[41]] Suppose one has conducted two measurements among dozens of students and prepared a scatter plot as shown in Fig. 9. The similarity between any two students, for example, student A and student B, is simply the Euclidian distance between the two dots. Clearly, there are two distinctive clusters in Fig. 9, one in the upper left corner and the other in the lower right corner. In determining the distance between the two clusters, one can use one of the following four measures:[42,43] distance between the closest points of two clusters (single linkage), distance between the farthest apart points of two clusters (complete linkage), average distance of all pairs between two clusters (average linkage), and distance between two mean vector locations (centroids, similar to the "center of mass" of a cluster of equal-mass particles) of two clusters. There is no superiority of one measure over another. Oftentimes, researchers have to try all of them to finally decide which one is the best for data interpretation.

The basic algorithm of cluster analysis involves iterations of assigning subjects into different clusters according to their distances to each cluster. Three methods are often used: agglomerative, divisive, and *K*-means.[43,44] The agglomerative method starts with individual subjects and considers each subject as a cluster of size 1. Two closest clusters first are merged to form a new cluster. This new cluster then is merged with the cluster (of size 1) nearest it.[45] This process continues until all subjects are combined into a single cluster. Consider a simple case of five data points in a two-dimensional space as shown in Fig. 10. First, the agglomerative method merges the two closest points *A* and *B* into a cluster (*AB*). Next, this cluster is merged with the nearest point *C* to form a new cluster (*ABC*). Then this new cluster is further merged with the nearer point *D* to form an even bigger cluster (*ABCD*). Finally, all points are included. An analogy of this method is adding leaves to branches and branches to trunks. The divisive method is just the reverse of the agglomerative method. It starts with one mother cluster that includes all subjects. Then it divides into two daughter clusters in a way that the distance between the daughter clusters is maximized. The dividing continues until every cluster contains a single subject. This method is similar to trunks diverging into branches and branches into leaves. *K*-means is different from the above two, and it does not form a tree. *K*-means requires a predetermined number (*K*) of clusters. Once the number *K* is specified, an initial partitioning is executed. Each subject then is assigned to a cluster whose centroid is closest. All subjects are either assigned or reassigned until no more movement is needed.

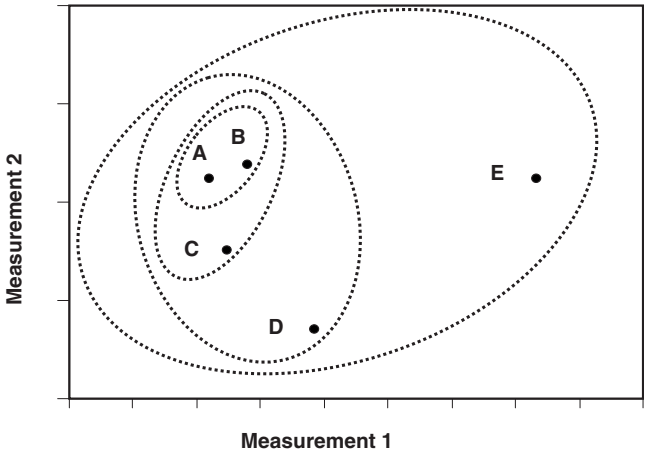


FIG. 10. An illustration of the agglomerative method.

We provide the following example to show how cluster analysis can be used for grouping students. Results are derived from a data set of 308 students who took the MIET. Five major topics are covered in this 33-item assessment: energy definition and representation, system specification, determination of work and heat, calculation of atomic spectra, and application of the work-energy theorem.[17] Our goal is to classify students into groups of distinctive performances on these five topics. We started with the students' scores on these five topics and used SAS to calculate Euclidian distances among the 308 students. SAS produces a $308 \times 308$ matrix, part of which is shown in Appendix A. We then used the agglomerative method and complete linkage to merge students. A tree plot of the procedure (also known as

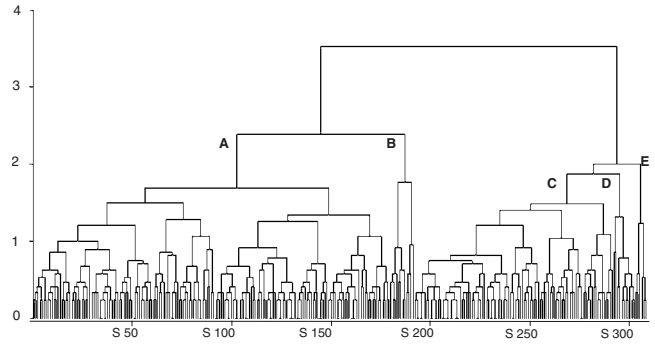


FIG. 11. A dendrogram for cluster analysis of 308 students. The horizontal axis represents 308 students (S1–S308); the vertical axis represents the Euclidian distance among clusters. For instance, the distance between clusters *A* and *B* is the vertical height of the horizontal line connecting *A* and *B*, which reads approximately 2.4.

dendrogram[46]) is depicted in Fig. 11. Here, the horizontal axis represents the arbitrarily assigned student identity and the vertical axis shows the Euclidian distance between clusters. As seen, each cluster extends a vertical line upward to meet and join with others. When two closest clusters merge, a horizontal line is placed to connect them. The height of the connecting horizontal line indicates the distance between the two clusters (see Fig. 11).

Several major clusters emerged in this example, as can be seen from the top down in Fig. 11. For a quick grasp of the results, we examine five clusters that are marked $A-E$ in Fig. 11. Here, we focus on five clusters mainly for illustration purpose. One certainly can choose a different number of clusters for analysis. For example, one can choose to study only the two biggest clusters (indicated by the top two vertical lines in Fig. 11) or as many as 308 clusters (corresponding to 308 students). Conceivably, results from too few or too many clusters are not as informative. Moreover, it is important to note that cluster analysis (like any other statistical technique) only reports *how* clusters are formed but not *why* they are formed as such. It is the researcher's job to find out the unique characteristics of the individuals in each cluster that have caused them to be grouped together. In this example, we seek to better understand what has made these five clusters different. Hence, we further construct Table V to examine their scores on each topic.

As seen from Table V, cluster *E* has the least number of students, but their scores on the "work-energy theorem" and "atomic spectra" are the highest. Cluster *D* has the next least number of students. Students in this cluster demonstrated highest scores on "energy definition," "system specification," and "determination of work and heat." Cluster *C* is similar to cluster *D* but with relatively lower scores. Thus, clusters *C* and *D* merge into one cluster that further joins with *E*. In other words, students in clusters *C* and *D* are comparable in the sense that they have fairly high scores on basic energy concepts and definitions, but their scores on "work-energy theorem" are significantly lower than those of group *E*. Students in cluster *E* seem to be at a higher level because they are not only able to grasp basic concepts but also able to successfully apply the work-energy theorem. As opposed to the above three clusters, clusters *A* and *B* both have the lowest scores on all topics. Nonetheless, cluster *A* differs from cluster *B* on the work-energy theorem and "work and heat" topics. This difference is easily noticed from the great vertical height of the horizontal line connecting *A* and *B* (see Fig. 11).

In the above example we used the agglomerative method to form clusters. We chose this method mainly because it generates more interpretable results than others. Another reason is that the agglomerative method is more efficient than the divisive and *K*-means methods. For the divisive method it is not feasible to optimize the initial partitioning when a data set is large since the number of ways to divide *N* subjects into two groups is $2^{N-1}-1$. As for the *K*-means procedure, the difficulty lies in the predetermination of the number of clusters. [Take the simplest situation for example. If we predetermine to divide students into two clusters on each of these five concepts (good or poor), then we will have $2^5=32$ clusters, more than what is manageable.] Nonetheless, both agglomerative and *K*-means have been adopted in recent physics educational studies. For example, a work by Springuel *et al.*[47] used agglomerative method for cluster analysis of student reasoning in kinematics; an unpublished work by Montenegro *et al.*[48] employed *K*-means to investigate student misconceptions in mechanics.

A final note on differences between cluster analysis and the aforementioned factor analysis is worth mentioning since both seem to perform data classifications. First, cluster analysis is often used to classify *subjects*, whereas factor analysis is employed to group *variables*. Consider the same example as before, where you administer a 30-item multiple-choice test among 300 students. If you are interested in examining how the 30 items interrelate with one another and

TABLE V. Five clusters and their scores (percentages) on different topics covered in the MIET. (Bold prints indicate the highest scores among the five clusters.)

| | Cluster | | | | |
|---|---|---|---|---|---|
| | *A* ($n=179$) | *B* ($n=12$) | *C* ($n=100$) | *D* ($n=11$) | *E* ($n=6$) |
| Energy definition and representation | 47 | 39 | 71 | **92** | 77 |
| System specification | 50 | 37 | 64 | **68** | 50 |
| Determination of work and heat | 43 | 32 | 56 | **64** | 50 |
| Calculation of atomic spectra | 68 | 67 | 82 | 82 | **92** |
| Application of work-energy theorem | 34 | 10 | 36 | 45 | **74** |

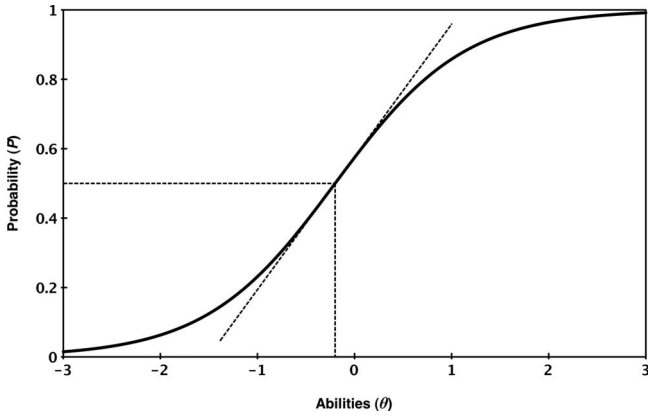


FIG. 12. An item characteristic curve.

thus reducing the number of variables, factor analysis is appropriate. On the other hand, if you want to know how student response patterns differ so as to classify students, cluster analysis should be considered. Second, cluster analysis often considers *Euclidian distances* for measuring similarity, not *correlations* as used in factor analysis. Why cannot one use correlations in cluster analysis? Consider the situation depicted in Fig. 9 for example, the two distinctive clusters are spatially apart but are highly correlated. Had correlations been used as a measure for similarity, subjects far apart in measurement space would have been placed into one cluster.

## V. ITEM RESPONSE THEORY

Item response theory (IRT) is a modern test theory. It can be used to estimate item characteristic parameters and examinees' latent abilities.[49] Here, item characteristic parameters include item difficulty and discrimination, which may seem the same as those in classical test theory but have different meanings and measures (see below). Examinees' latent abilities are referred to as examinees' general knowledge, capabilities, and skills in a specific domain. IRT assumes one "unidimensional" skill or ability that underlies examinees' responses to all items. This skill or ability is considered as latent because it is a nonphysical entity and is not directly measured. For example, a student's score on an E&M multiple-choice test is only an outcome of her understanding of E&M but is not her understanding itself. Simply put, an E&M test cannot measure students' understanding the same way as a ruler can measure a table length. IRT, however, intends to provide an estimate of such unmeasurable entities.

The basic task of IRT is to use logistic regression to formulate observed binary data. A graphical representation of this logistic regression (also known as the item characteristic curve[50]) is depicted in Fig. 12. Here, the horizontal axis represents latent ability $\theta$ and the vertical axis shows the probability $P(\theta)$ of answering an item correctly. Two parameters are useful in describing the shape of the curve. One is the location of the curve's middle point; the other is the slope of the curve at the middle point. The middle point is at $P(\theta)=0.5$, and its corresponding value along the ability scale $\theta$ is defined as item difficulty. In other words, item difficulty is the ability value at a 50% probability of correct response. So, the greater the difficulty value of a particular test item, the higher an ability level is required to have a 50% probability of correct response. This is different from the difficulty measure in classical test theory (Sec. II). As for the slope of the curve, it has a maximum at the middle point. If the slope is large, the curve is steeper, indicating that students of high abilities have a greater probability of correct response than those of low abilities. Conversely, if the middle point slope is small, the curve is flatter; students of high abilities have nearly the same probability of correct response as those of low abilities. In this sense, the slope at the middle point is a measure of an item's discrimination. In IRT, the item difficulty and discrimination parameters are denoted by $b$ and $a$, respectively. Using these notions, the mathematical expression of this logistic model is given by

$$P(\theta)=\frac{1}{1+e^{-a(\theta-b)}}.$$

The core of IRT is to determine the item characteristic parameters $a$ and $b$ and the students' latent abilities $\theta$.

Three IRT models based on the logistic regression are frequently used; they are the one-parameter Rasch model, a two-parameter model, and the three-parameter Birnbaum model.[51] The two-parameter model is identical to the above logistic regression with both parameters $a$ and $b$ undetermined. For the one-parameter Rasch model, item discrimination $a$ is held constant ($a=1$) for all items, thus leaving the model with only one undetermined parameter, which is $b$. The three-parameter Birnbaum model considers a guessing effect and introduces a new parameter $c$. This parameter $c$ represents the probability of guessing a correct response for those who do not possess the necessary ability to answer it correctly. Thus, the observed probability of correct response now becomes

$$c[1-P(\theta)]+P(\theta)=c+(1-c)\frac{1}{1+e^{-a(\theta-b)}}.$$

In this model, three parameters need to be determined.

To demonstrate how IRT can be a useful tool in evaluating multiple-choice items, we provide the following example using the three-parameter Birnbaum model calculated via MULTILOG.[52] Results are based on binary data collected from 308 students' responses to 33 items in the MIET. Recall that our goal is to estimate the item characteristic parameters $a$, $b$, and $c$ for each of the individual items. For illustration purposes, we show in Fig. 13 item characteristic curves of two items: Q27 and Q33.

As seen, item 27 has a positive discrimination value $a=0.74$, displaying a monotonically increasing "S" curve in the range of $\theta\in[-3,+3]$. The value along the $\theta$ scale for the curve middle point is 1.25, meaning its item difficulty is $b=1.25$. Simply put, students whose ability value is 1.25 have a 50% probability of correctly answering this item. The lower left part of the curve shows an asymptote of 0.13, indicating that students of low abilities have a 13% probability of guessing this item correctly. As opposed to item 27, item 33 displays nearly a flat line ($a=0.04$) in the $\theta\in[-3,+3]$ range, indicating the item fails to distinguish students of

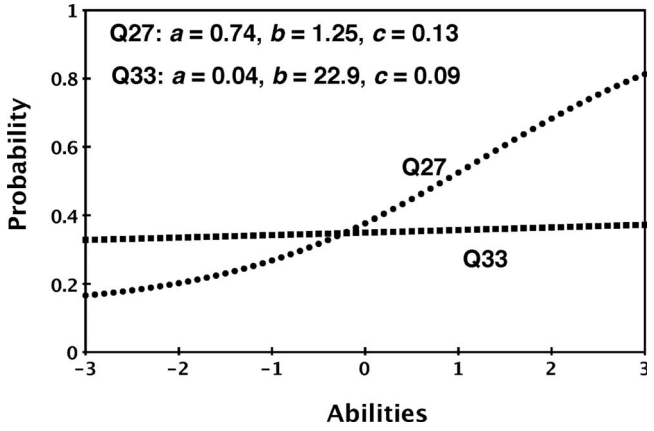


FIG. 13. Item characteristic curves for two MIET items (Q27 and Q33).

high abilities from those of low abilities. The reason for this may be due to the high difficulty level of this item ($b=22.9$). In future revisions, we will consider modifying item 33 to make it easier.

In the above example, the ability scale can be generally described as students' knowledge of energy topics in the M&I mechanics course. The reason is twofold. First, IRT assumes a unidimensional scale for the entire test. Second, the MIET solely focuses on energy topics that are covered in the M&I mechanics course. Of note here is that IRT-estimated abilities may be correlated with, but are not identical to, test total scores. A total score may be dependent on the specific questions used in a test, whereas IRT-estimated abilities are independent of the questions used in a test. For an elaborated proof, refer to Ref. 50. Similarly, the item difficulty and discrimination parameters in IRT are also independent of examinees who take the test.

In addition to the above application, IRT can be used to evaluate the functions of distracters in each item. The basic idea is to examine trace lines for alternative choices. As an example, we plot in Fig. 14 alternative-choice trace lines for one item (Q30) in the MIET using the Bock-Samejima model[53,54] (see Appendix B for this question). In this example, the correct choice (choice 2) displays a monotonically

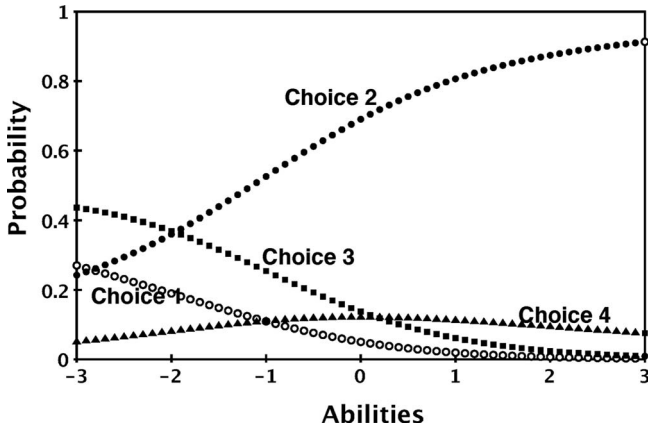


FIG. 14. Trace lines for alternative choices of one MIET item (Q30).

increasing S curve in the $\theta \in [-3,+3]$ range. Therefore, students of high abilities are more likely to choose the correct answer than those of low abilities. As for choices 1 and 3, the trace lines have a reverse trend. So, students of low abilities are more likely to select choice 1 or 3 than those of high abilities. Take choice 3 for example; the probability of choosing this answer is less than 1% for students of an ability value of +3, but it is as high as 44% for those of an ability value of −3. As for choice 4, the trace line is relatively flat and low, suggesting that not many students choose this answer at any ability level. Therefore, alternative choices 1 and 3 seem to function better than choice 4 in distracting students of low abilities.

In Fig. 14, we once again plot probabilities against latent abilities, not test total scores. In fact, it is much easier to use total scores as a substitute for abilities than to perform IRT. This is particularly true when one lacks adequate knowledge of IRT or does not have computer software. As a rudimentary step toward IRT, using total scores as a substitute for $\theta$ can provide a glimpse of what real IRT curves may look like. A recent study by Morris *et al.* used this rudimentary approach to evaluate Force Concept Inventory items.[55] Conversely, a paper by Lee *et al.* employed two-parameter IRT to measure student latent abilities in answering Newtonian physics questions in a web-based physics tutoring system MASTERINGPHYSICS.[56]

Finally, some practical issues of IRT are worth noting. First, a large sample size generally is recommended for a good model fit. Since the Rasch model estimates fewest parameters, a data set of "as few as 100" may be needed for stable results.[57] (Linacre[58] suggested 50 for the simplest Rasch model.) For other models, a sample size of several hundred often is required. Also, different study purposes may call for different sample sizes. For example, calibration of high-stake test items may require sample sizes over 500 to ensure accuracy. But for low-stake tests, "one does not need large sample sizes."[57] Second, prior to IRT one may consider performing factor analysis to determine if there is a single prominent factor. The reason for doing so lies in the IRT assumption that there is one single factor underlying examinees' responses. If factor analysis yields a single prominent factor, one then can proceed with IRT. If factor analysis yields multiple factors, one then can use IRT within each factor. In fact, recent development in IRT has relaxed such a constraint on "unidimensionality," thus allowing researchers to perform IRT even for data with multidimensional factors.[59] Third, there have been great controversies on IRT model selections. Though the three-parameter model seems to be the most complicated and hence the most stringent model, arguments have been made that it in fact is the most general model and that the other two models (the Rasch and two-parameter models) are just special cases of the three-parameter model with constraints on $a$'s and $c$'s.[60] Therefore, it is recommended that the three-parameter should be used as a start.[60] On the other hand, the seemingly simple expression of the Rasch model continues to attract many researchers. In either case, one may rely on IRT software to examine whether a model provides a good fit and then decide which model to choose.[52,61] Thorough discussions on these issues are beyond the scope of the paper. Interested readers can refer to Refs. 60,61 for more information.

## VI. MODEL ANALYSIS

Model analysis[62] is a PER-initiated approach to assessing student learning dynamics. The goal of model analysis is to present the probabilities of students' use of different "models" in answering isomorphic questions that test the same underlying concepts but have different surface features. Here, models are purposefully defined rather vaguely to include both facets (bits and pieces of irreducible information) and locally (or universally) coherent knowledge resources.[63] For example, when encountering a question where only the work-energy theorem is applicable, a student may approach the question by using any of the following knowledge or information: the work-energy theorem, the impulse-momentum theorem, or some completely irrelevant formula. In this example, each approach is considered a model. The first approach is regarded as a "correct" model, the second can be conveniently labeled as "impulse-momentum" model, and the third can be named as "others."

The basic algorithm of model analysis starts with a linear vector $\boldsymbol{Q}$ that represents an individual's probabilities of using different models in answering a set of questions. For example, a student answers four questions, all of which require an application of the work-energy theorem. Suppose the student correctly applies the work-energy theorem in only two questions but uses the impulse-momentum theorem and some irrelevant formula in the other two questions, respectively. Then the vector $\boldsymbol{Q}$ for this student is

$$\boldsymbol{Q} = 0.50\begin{pmatrix}1\\0\\0\end{pmatrix} + 0.25\begin{pmatrix}0\\1\\0\end{pmatrix} + 0.25\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}0.50\\0.25\\0.25\end{pmatrix}.$$

Here, the three elements "0.50," "0.25," and "0.25" indicate the frequencies of applying three different models, respectively. Use the square root of each element in $\boldsymbol{Q}$ to form a new vector $\boldsymbol{V}$,

$$\boldsymbol{V} = \begin{pmatrix}\sqrt{0.50}\\\sqrt{0.25}\\\sqrt{0.25}\end{pmatrix} = \begin{pmatrix}0.71\\0.50\\0.50\end{pmatrix}.$$

Now take an outer product of $\boldsymbol{V}$ with a transpose of itself $\boldsymbol{V}\otimes\boldsymbol{V}^{\mathbf{T}}$ to get a matrix, namely, the "density matrix" for each individual student. Next, take an average over all the students to obtain a class density matrix. Depending on how students use different models, the class density matrix may display different patterns. Figure 15 shows three examples as discussed by Bao and Redish.[63] The first matrix has only one nonzero element along its diagonal, meaning the entire class uses the same model consistently. The second matrix has three nonzero elements along its diagonal, indicating that although the entire class has three models, each student uses only one model consistently. The third matrix in Fig. 15 shows a mixed case where students use three models inconsistently. Now solve the class density matrix for its eigenvalues and eigenvectors. The eigenvector of the largest eigenvalue represents the dominant model used by the class.

$$\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix} \quad \begin{pmatrix}0.5&0&0\\0&0.3&0\\0&0&0.2\end{pmatrix} \quad \begin{pmatrix}0.5&0.2&0.1\\0.2&0.3&0.1\\0.1&0.1&0.2\end{pmatrix}$$

**(a)** **(b)** **(c)**

FIG. 15. Examples of class density matrix as appeared in Ref. 54: (a) the entire class uses model 1 consistently, (b) the class consists of three groups of students each with a consistent model, and (c) students use multiple models inconsistently.

**A point particle of mass *m* is initially at rest. A constant force is applied to the point particle during a first time interval, while the particle's speed increases from 0 to *v*. The constant force continues to act on the particle for a second time interval, while the particle's speed increases further from *v* to *2v*. Consider the work done on the particle. During which time interval is more work done on the point particle?**

**(a) The first time interval**
**(b) The second time interval**
**(c) The work done during the two time intervals is the same**
**(d) Not enough information to determine**

FIG. 16. One item (Q33) from the MIET.

In the following example, we use model analysis to study students' responses to four questions in the MIET that require an application of the work-energy theorem. Results obtained herein are derived from 300 students who answered all four questions. (Eight students with missing responses are excluded from the analysis.[64]) We consider three models for analysis: (1) work-energy model, (2) momentum or force-motion model, and (3) others. In Fig. 16 we display one item (Q33) to illustrate what each model may look like. In this item, choice (b) is the correct answer, thus representing the "work-energy" model. Many students answered (c) and argued during interviews that the work done is the same in

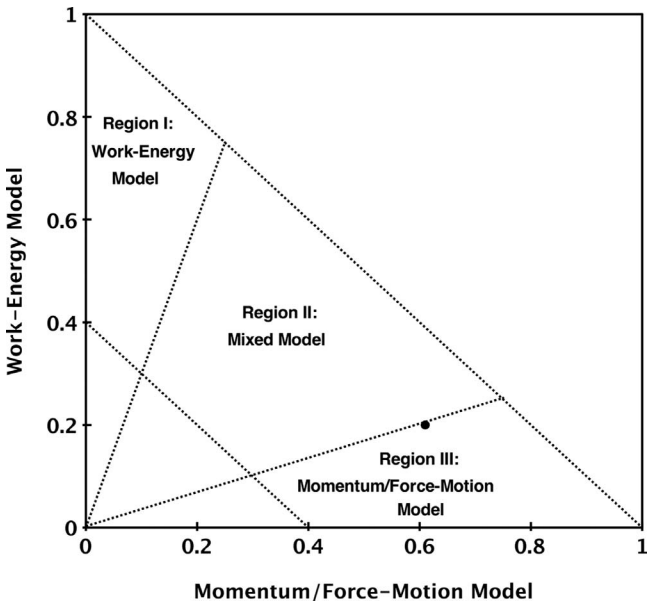


FIG. 17. A model plot of students' answering four MIET items.

both time intervals because the change in speed is the same. Apparently, these students focused on motion change instead of energy change; therefore, their model can be described as the "momentum or force-motion" model. As for choices (a) and (d), we label them as others.

We obtain a class density matrix, as shown below, by averaging individual density matrices,

$$\begin{pmatrix} 0.275 & 0.304 & 0.06 \\ 0.304 & 0.638 & 0.121 \\ 0.06 & 0.121 & 0.085 \end{pmatrix}.$$

We then solve its eigenequation and find that the largest eigenvalue is $\lambda_{max}=0.835$ and its corresponding eigenvector is $\boldsymbol{U}=(-0.484,-0.857,-0.177)^{\mathbf{T}}$. According to Bao and Redish,[63] the class probability of using the first model (work-energy model) can thus be calculated as $\lambda_{max}\times \boldsymbol{U}_1^2=0.835\times(-0.484)^2=0.20$. Similarly, the class probability of using the momentum or force-motion model is 0.61. As for the third model others, the probability is $\lambda_{max}\times \boldsymbol{U}_1^2=0.835\times(-0.177)^2=0.03$, which is negligible. Note that the sum of the above three probabilities is less than 1, implying that these probabilities do not form a complete set. In fact, the probabilities originating from the second and the third eigenvectors have not been taken into account. Bao and Redish considered these additional eigenvectors as "corrections of less popular features that are not presented by the primary state."[63] Because of these "corrections," probabilities from model analysis are lower than what is calculated from a direct division of the number of students choosing a particular model by the total number of instances. This is generally true except for some extreme cases in which, for example, all students consistently use the same model. Because model analysis takes into account the kind of inconsistencies, a class density matrix almost always displays nonzero off-diagonal elements.

Here, we represent the first two probabilities in a "model plot" (Fig. 17) proposed by Bao and Redish, in which three regions are separated by two straight lines passing through the origin with slopes of 3 and 1/3, respectively.[63] The data point falls into the region of the "momentum or motion-force" model but is close to the borderline. We thus conclude that students are likely to use both the work-energy and momentum or force-motion models although the latter is more dominant.

For researchers who consider using model analysis, it is worth noting some requirements this analysis places on users. First, one needs to have in a test "a sequence of questions or situations in which an expert would use a single coherent mental model."[63] Simply put, a sequence of questions that target the same concept or principle is necessary. In fact, a large number of such questions is strongly preferred because "the probabilistic character of the student model state arises from the presentation of a large number of questions or scenarios."[63] Second, these questions ought to have distracters that represent similar models. These models must be predetermined (both identified and categorized) before performing model analysis. In other words, qualitative studies of students' naive conceptions and incorporation of results into item distracters are a necessary precursor to model analysis.

## VII. SUMMARY AND DISCUSSION

In this paper, we discuss the goals, basic algorithms, and applications of five approaches to analyzing multiple-choice test data; they are classical test theory, factor analysis, cluster analysis, item response theory, and model analysis. These approaches are valuable tools for PER studies on large-scale measurement using multiple-choice assessments.

Among them, classical test theory and item response theory are particularly useful for test evaluations during the development stage of an assessment. Both approaches can yield detailed information on difficulty and discrimination of individual items and thus are regularly employed to identify problematic items. Classical test theory also gives rise to the measures of an entire test, revealing the overall reliability and discrimination of the test. (This is why it is called "test theory.") Owing to its simple theoretical framework, classical test theory has rather straightforward formulations. Thus, it is easy to carry out. Nevertheless, classical test theory has some major weaknesses. A prominent one is that results on item measures are examinee dependent, and examinees' total scores are item dependent. Conversely, item response theory overcomes such weaknesses by assuming a single underlying "ability" (or "skill") and using logistic regression to describe the propensity of correct responses to individual items. As a result, its estimated item measures and examinee abilities are mutually independent. Moreover, item response theory can also be used to examine how alternative choices function. (Because of its emphasis on individual items, this theory is named "item response theory.")

Factor analysis and cluster analysis are good candidates for grouping data into different categories. However, the former intends to group variables (test items) into a small number of components (principle component analysis) or factors (common factor analysis), whereas the latter intends to group subjects (test examinees) into different clusters. Hence, factor analysis reveals information on how test items are interrelated, while cluster analysis illuminates how students' responses differ. A noteworthy aspect about factor analysis is the subtle difference between principle component analysis and common factor analysis. Principle component analysis is a pure data reduction technique and does not presume underlying "causal structures." But common factor analysis assumes some common factors to be the cause of observed data.

Model analysis is a useful tool for studies on how consistently students answer isomorphic questions. It utilizes quantum notations and eigenvalue analysis techniques to

examine the probabilities of students' using different mental models in answering isomorphic questions (given such questions are readily available). Because model analysis requires that all questions have alternative choices covering the same mental models, qualitative studies must be performed first to ensure these questions meet the needs.

All the above approaches are powerful tools for data analysis of multiple-choice tests. However, none is perfect or can be exclusively used to answer all research questions. As discussed before, each has its specific purposes and applications. Hence, caution must be practiced when selecting among these approaches. Finally, it is important to note that our ultimate goal is to make sense of raw data. Therefore, when encountering two (or more) equally sound choices, one should always prefer the one that better (or best) facilitates data interpretation.

TABLE VI. This table shows detailed item analysis results for MIET based on 389 data collected from two institutions: North Carolina State University and Purdue University. The average total score is 17.4 out of 33 and the standard deviation is 4.9.

| Item | Difficulty index | Discrimination index | Point biserial coefficient |
|---|---|---|---|
| Q1 | 0.24 | 0.4 | 0.39 |
| Q2 | 0.65 | 0.57 | 0.46 |
| Q3 | 0.39 | 0.64 | 0.52 |
| Q4 | 0.88 | 0.22 | 0.24 |
| Q5 | 0.76 | 0.25 | 0.25 |
| Q6 | 0.79 | 0.14 | 0.18 |
| Q7 | 0.31 | 0.57 | 0.50 |
| Q8 | 0.26 | 0.41 | 0.37 |
| Q9 | 0.82 | 0.28 | 0.26 |
| Q10 | 0.52 | 0.45 | 0.37 |
| Q11 | 0.83 | 0.27 | 0.26 |
| Q12 | 0.72 | 0.5 | 0.39 |
| Q13 | 0.65 | 0.59 | 0.42 |
| Q14 | 0.79 | 0.42 | 0.32 |
| Q15 | 0.43 | 0.62 | 0.47 |
| Q16 | 0.32 | 0.47 | 0.38 |
| Q17 | 0.73 | 0.29 | 0.23 |
| Q18 | 0.34 | 0.39 | 0.35 |
| Q19 | 0.65 | 0.30 | 0.30 |
| Q20 | 0.66 | 0.27 | 0.25 |
| Q21 | 0.33 | 0.3 | 0.26 |
| Q22 | 0.37 | 0.39 | 0.27 |
| Q23 | 0.43 | 0.41 | 0.31 |
| Q24 | 0.13 | 0.10 | 0.15 |
| Q25 | 0.80 | 0.37 | 0.30 |
| Q26 | 0.59 | 0.26 | 0.16 |
| Q27 | 0.39 | 0.56 | 0.47 |
| Q28 | 0.36 | 0.18 | 0.18 |
| Q29 | 0.67 | 0.41 | 0.34 |
| Q30 | 0.67 | 0.45 | 0.36 |
| Q31 | 0.33 | 0.32 | 0.28 |
| Q32 | 0.29 | 0.54 | 0.43 |
| Q33 | 0.32 | 0.33 | 0.29 |


## ACKNOWLEDGMENTS

The authors would like to thank Richard Hake for his insightful comments on the paper. The authors also greatly appreciate the support from the Physics Education Research Groups at The Ohio State University and North Carolina State University.


## APPENDIX A: DATA

Tables VI and VII show item analysis results for MIET and student score distribution.

The following is a correlation matrix $\boldsymbol{R}$ for eight items (Q12–Q14, Q16, Q27, and Q31–Q33) selected from the M&I Energy Test (MIET). This matrix is symmetric with respect to its diagonal ($R_{ij}=R_{ji}$); the upper half of the matrix is thus omitted. We use this matrix $\boldsymbol{R}$ for principal component analysis of these eight items,

TABLE VII. This table shows student score distribution.

| Total score | No. of students | Total score | No. of students |
|---|---|---|---|
| 0 | 0 | 17 | 39 |
| 1 | 0 | 18 | 27 |
| 2 | 0 | 19 | 25 |
| 3 | 0 | 20 | 17 |
| 4 | 1 | 21 | 25 |
| 5 | 0 | 22 | 14 |
| 6 | 1 | 23 | 14 |
| 7 | 2 | 24 | 15 |
| 8 | 2 | 25 | 10 |
| 9 | 4 | 26 | 3 |
| 10 | 13 | 27 | 7 |
| 11 | 17 | 28 | 6 |
| 12 | 17 | 29 | 4 |
| 13 | 25 | 30 | 3 |
| 14 | 39 | 31 | 0 |
| 15 | 29 | 32 | 1 |
| 16 | 29 | 33 | 0 |

$$\boldsymbol{R} = \begin{pmatrix} 1 & . & . & . & . & . & . & . \\ 0.38537 & 1 & . & . & . & . & . & . \\ 0.20388 & 0.20578 & 1 & . & . & . & . & . \\ 0.11881 & 0.21611 & 0.20957 & 1 & . & . & . & . \\ 0.09215 & 0.13006 & 0.03928 & 0.13247 & 1 & . & . & . \\ 0.09739 & 0.03541 & -0.00308 & 0.03355 & 0.16152 & 1 & . & . \\ 0.15976 & 0.12583 & 0.06139 & 0.11005 & 0.13566 & 0.14278 & 1 & . \\ 0.06961 & 0.07763 & 0.06142 & 0.03742 & 0.17751 & 0.17413 & 0.25602 & 1 \end{pmatrix}.$$

The following is an adjusted correlation matrix $\boldsymbol{R}_{\text{adjust}}$ for the aforementioned eight items. Of note is that the diagonal elements are no longer 1's. Instead, they are replaced with SAS estimated variances that are accounted for by the common factors. We use the adjusted correlation matrix $\boldsymbol{R}_{\text{adjust}}$ for common factor analysis of eight MIET items,

$$\boldsymbol{R}_{\text{adjust}} = \begin{pmatrix} 0.182 & . & . & . & . & . & . & . \\ 0.385 & 0.196 & . & . & . & . & . & . \\ 0.204 & 0.206 & 0.090 & . & . & . & . & . \\ 0.119 & 0.216 & 0.210 & 0.091 & . & . & . & . \\ 0.092 & 0.130 & 0.039 & 0.132 & 0.076 & . & . & . \\ 0.097 & 0.035 & -0.003 & 0.034 & 0.162 & 0.062 & . & . \\ 0.160 & 0.126 & 0.061 & 0.110 & 0.136 & 0.143 & 0.105 & . \\ 0.070 & 0.078 & 0.061 & 0.037 & 0.178 & 0.174 & 0.256 & 0.103 \end{pmatrix}.$$

The following is part of the $308 \times 308$ similarity matrix $\boldsymbol{D}$ for 308 students who took the MIET. Each student forms a response vector $(T_1, T_2, T_3, T_4, T_5)$, where $T_i$ represents his or her score on the $i$th topic. The scores on each topic display both order and magnitude and thus can be used as interval data.[65] The similarities among students are simply calculated as the Euclidian distances between their response vectors. Therefore, element $D_{ij}$ in the following matrix represents the Euclidian distance between the response vector of student $i$ and that of student $j$. Due to limited space, we only show the Euclidian distances among ten students. Because this matrix is symmetric with respect to its diagonal, we omit its upper half. This matrix $\boldsymbol{D}$ is used to perform agglomerative cluster analysis,

$$\boldsymbol{D} = \begin{pmatrix} 0 & . & . & . & . & . & . & . & . & . \\ 5.57 & 0 & . & . & . & . & . & . & . & . \\ 5.00 & 3.16 & 0 & . & . & . & . & . & . & . \\ 4.90 & 2.65 & 3.61 & 0 & . & . & . & . & . & . \\ 6.93 & 1.73 & 4.58 & 3.46 & 0 & . & . & . & . & . \\ 6.71 & 2.00 & 4.69 & 3.87 & 1.00 & 0 & . & . & . & . \\ 5.00 & 5.29 & 4.24 & 6.25 & 6.40 & 5.83 & 0 & . & . & . \\ 9.33 & 5.83 & 6.32 & 7.81 & 5.92 & 5.48 & 5.29 & 0 & . & . \\ 3.74 & 4.36 & 2.24 & 3.46 & 5.83 & 5.92 & 4.58 & 7.81 & 0 & . \\ 9.75 & 4.47 & 7.07 & 6.56 & 3.32 & 3.46 & 8.25 & 5.48 & 8.66 & 0 \end{pmatrix}.$$

## APPENDIX B: SELECTED QUESTIONS FROM MIET

Consider a system that consists of two asteroids in deep space. The following figure plots energy of several different states of the two-asteroid system versus distance $r$ between them. Questions 12–14 refer to this figure (see Fig. 18).

Q12. What does the curved line I represent?
(a) Kinetic energy
(b) Potential energy
(c) Sum of kinetic energy and potential energy
(d) Rest energy
(e) None of the above

Q13. What do the horizontal lines (II, III, and IV) represent?
(a) Kinetic energy
(b) Potential energy
(c) Sum of kinetic energy and potential energy
(d) Rest energy

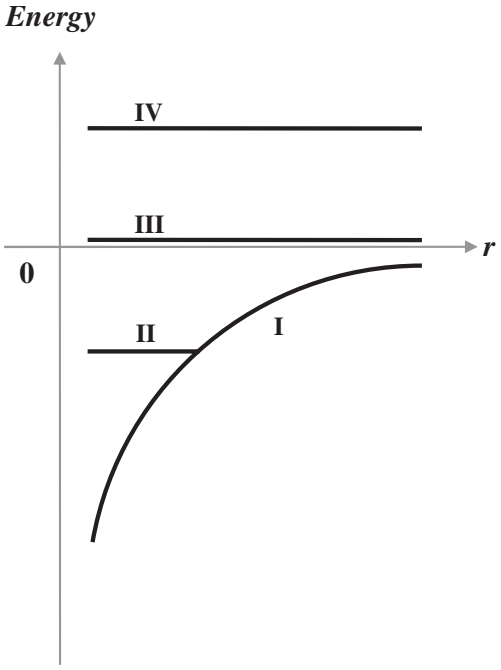


FIG. 18. Energy plot of several different states of the two-asteroid system vs the distance $r$ between them.

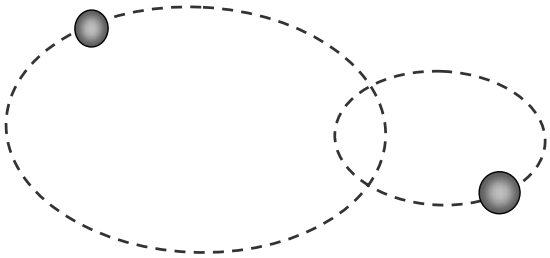

FIG. 19. Diagram of two stars in deep space orbiting around each other.

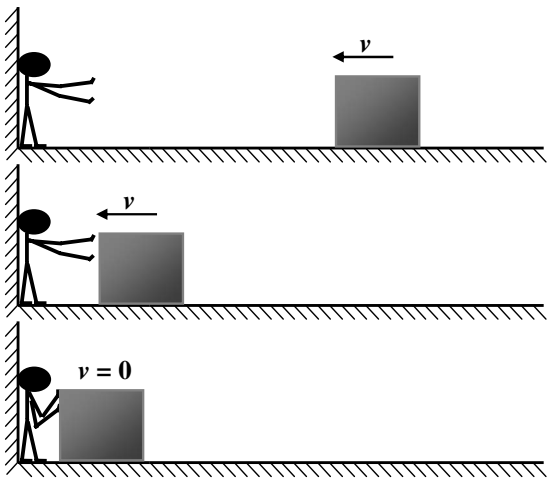


FIG. 20. Diagram of a box of mass $M$ moving at speed $v$ toward a person along a surface of negligible friction.

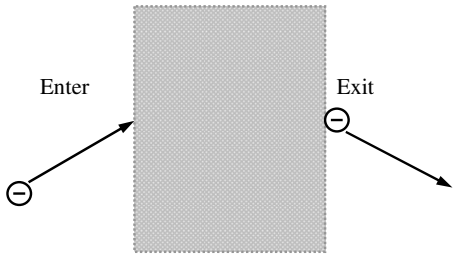


FIG. 21. Diagram of an electron entering and exiting a box.

(e) None of the above

Q14. Which of the horizontal lines represents a bound state?

(a) Line II
(b) Line III
(c) Line IV
(d) All of the above
(e) None of the above

Q16. In deep space two stars are orbiting around each other. Consider the sum of kinetic energy and gravitational potential energy $K+U$ of the two-star system. Which of the following statements is true? (See Fig. 19.)

(a) $K+U$ is positive
(b) $K+U$ is zero
(c) $K+U$ is negative
(d) $K+U$ is either positive or zero
(e) $K+U$ is either negative or zero

Q27. A box of mass $M$ is moving at speed $v$ toward a person along a surface of negligible friction. A person leans against a wall with both arms stretched and applies a pushing force to stop the box. Finally the person brings the box to a full stop. There is no temperature change in the box at any time (see Fig. 20). We can conclude that during the process:

(a) The amount of work done on the box by the person was $+\frac{1}{2}Mv^2$
(b) The amount of work done on the box by the person was positive, but there is not enough information to determine the value
(c) The amount of work done on the box by the person was 0
(d) The amount of work done on the box by the person was $-\frac{1}{2}Mv^2$

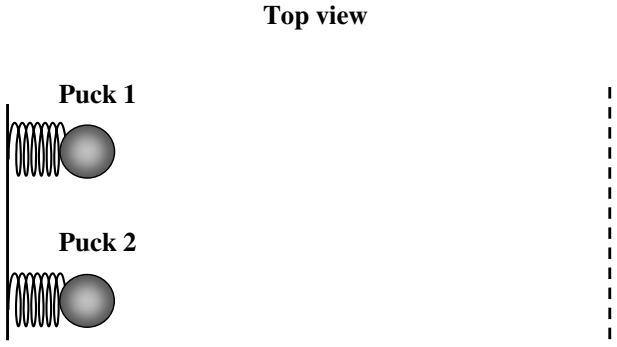


FIG. 22. Diagram of a low-friction table with two pucks launched by pushing them against two identical springs.

(e) The amount of work done on the box by the person was negative, but there is not enough information to determine the value

Q30. A clay ball moving to the right at a certain speed has kinetic energy of 10 J. An identical clay ball is moving to the left at the same speed. The two balls smash into each other and both come to a stop. What happened to the energy of the two clay-ball system?

(a) The kinetic energy of the system did not change

(b) The kinetic energy changed into thermal energy

(c) The total energy of the system decreased by an amount 20 J

(d) The initial kinetic energy of the system was zero, so there was no change in energy

Q31. An electron with speed $1\times10^8$ m/s enters a box along a direction depicted in the diagram. Some time later the electron is observed leaving the box with the same speed of $1\times10^8$ m/s but different direction as shown (see Fig. 21). From this, we can conclude that in the box:

(a) The net force on the electron was nonzero, and the net work done on the electron was nonzero

(b) The net force on the electron was nonzero, but the net work done on the electron was zero

(c) The net force on the electron was zero, but the net work done on the electron was nonzero

(d) The net force on the electron was zero, and the net work done on the electron was zero

(e) Not enough information to determine

Q32. On a low-friction table you launch two pucks by pushing them against two identical springs through the same amount. The two pucks have the same shape and size, but the mass of puck 2 is twice the mass of puck 1. Then you release the pucks and the springs propel them toward the finish line. At the finish line, how does the kinetic energy of puck 1 compare to the kinetic energy of puck 2? (See Fig. 22.)

(a) It is the same as the kinetic energy of puck 2

(b) It is twice the kinetic energy of puck 2

(c) It is half the kinetic energy of puck 2

(d) It is four times the kinetic energy of puck 2

(e) It is one-forth the kinetic energy of puck 2

Q33. A point particle of mass $m$ is initially at rest. A constant force is applied to the point particle during a first time interval, while the particle's speed increases from 0 to $v$. The constant force continues to act on the particle for a second time interval, while the particle's speed increases further from $v$ to $2v$. Consider the work done on the particle. During which time interval is more work done on the point particle?

(a) The first time interval

(b) The second time interval

(c) The work done during the two time intervals is the same

(d) Not enough information to determine


[1] See, for example, A. Bork, Letters to the editor, Am. J. Phys. **52**, 873 (1984); R. Varney, More remarks on multiple choice questions, *ibid.* **52**, 1069 (1984); T. Sandin, On not choosing multiple choice, *ibid.* **53**, 299 (1985); M. Wilson and M. Bertenthal, *Systems for State Science Assessment* (National Academies Press, Washington, DC, 2006).

[2] See, for example, R. Lissitz and K. Samuelsen, A suggested change in terminology and emphasis regarding validity and education, Educ. Res. **36**, 437 (2007); S. Ramlo, Validity and reliability of the force and motion conceptual evaluation, Am. J. Phys. **76**, 882 (2008).

[3] L. Cronbach and L. Furby, How we should measure "change": Or should we?, Psychol. Bull. **74**, 68 (1970); L. Suskie, *Assessing Student Learning: A Common Sense Guide*, 2nd ed. (Jossey-Bass, San Francisco, 2009); R. R. Hake, Can distance and classroom learning be increased?, J. Scholarship Teach. Learn. **2**, 1 (2008); http://www.physics.indiana.edu/~hake/MeasChangeS.pdf

[4] T. Kline, *Psychological Testing: A Practical Approach to Design and Evaluation* (SAGE, Thousand Oaks, CA, 2005), pp. 91–105.

[5] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980) (http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/38/90/26.pdf).

[6] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 97.

[7] A. Oosterhof, *Classroom Applications of Educational Measurement*, 3rd ed. (Merrill, Upper Saddle River, NJ, 2001) pp. 176–178.

[8] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 99.

[9] E. Ghiselli, J. Campbell, and S. Zedeck, *Measurement Theory for the Behavioral Sciences* (Freeman, San Francisco, 1981).

[10] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, London, 1986), p. 143.

[11] G. Kuder and M. Richardson, The theory of the estimation of test reliability, Psychometrika **2**, 151 (1937).

[12] J. Bruning and B. Kintz, *Computational Handbook of Statistics*, 4th ed. (Longman, New York, 1997).

[13] R. Doran, *Basic Measurement and Evaluation of Science Instruction* (NSTA, Washington, DC, 1980), p. 104.

[14] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, London, 1986), p. 150.

[15] L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 010105 (2006).

[16] P. Kline, *A Handbook of Test Construction: Introduction to Psychometric Design* (Methuen, London, 1986), p. 144.

[17] L. Ding, Ph.D. thesis, North Carolina State University, 2007 (http://www.lib.ncsu.edu/theses/available/etd-06032007-181559/).

[18] R. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

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

[20] C. Singh and D. Rosengrant, Multiple-choice test of energy and

momentum concepts, Am. J. Phys. **71**, 607 (2003).
[21] P. Engelhardt and R. Beichner, Students' understanding of direct current resistive electric circuits, Am. J. Phys. **72**, 98 (2004).
[22] L. Hatcher, *A Step-By-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling* (SAS Institute, Cary, NC, 1994).
[23] R. Gorsuch, Common factor analysis versus component analysis: Some well and little known facts, Multivar. Behav. Res. **25**, 33 (1990).
[24] R. Nichols, *Multiple Analysis Program System* (MAPS, Amherst, NY, 1985), pp. 13–15.
[25] R. Gorsuch, *Factor Analysis* (Lawrence Erlbaum, Hillsdale, NJ, 1983), pp. 165–169.
[26] I. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 2002), pp. 269–298; G. Dunteman, *Principal Component Analysis* (SAGE, Newbury Park, CA, 1989) pp. 48–49.
[27] R. Gorsuch, *Factor Analysis* (Lawrence Erlbaum, Hillsdale, NJ, 1983), pp. 67–71 and 175–238.
[28] J. Hair, R. Anderson, R. Tatham, and W. Black, *Multivariate Data Analysis with Readings*, 5th ed. (Prentice-Hall, Englewood Cliffs, NJ, 1998) p. 111.
[29] SAS/STAT, Version 8, Cary, NC, 2000.
[30] J. Kim and C. Mueller, *Introduction to Factor Analysis: What It Is and How to Do It* (SAGE, Beverly Hills, CA, 1978).
[31] J. Kim and C. Mueller, *Factor Analysis: Statistical Methods and Practical Issues* (SAGE, Beverly Hills, CA, 1978).
[32] L. Hatcher, *A Step-By-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling* (SAS Institute, Cary, NC, 1994), p. 71.
[33] J. Loehlin, Component analysis versus common factor analysis: A case of disputed authorship, Multivar. Behav. Res. **25**, 29 (1990).
[34] C. Spearman, General intelligence, objectively determined and measured, Am. J. Psychol. **15**, 201 (1904).
[35] For example, consider two items that both are scored either "correct" (1) or "incorrect" (0). Thus, possible outcomes are (1, 1), (1, 0), (0, 1), and (0, 0). If $a$, $b$, $c$, and $d$ are used to indicate the frequencies of the four outcomes, respectively, the correlation between the two dichotomous variables then can be calculated as $\boldsymbol{ad-bc}/\sqrt{(\boldsymbol{a+b})(\boldsymbol{c+d})(\boldsymbol{a+c})(\boldsymbol{b+d})}$.
[36] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[37] W. Adams, K. Perkins, N. Podolesfsky, M. Dubson, N. Finkelstein, and C. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).
[38] B. Everitt, *Cluster Analysis* (Heinemann, London, 1974).
[39] B. Everitt, *Cluster Analysis* (Heinemann, London, 1974), p. 40.
[40] M. Aldenderfer and R. Blashfield, *Cluster Analysis* (SAGE, Beverly Hills, CA, 1984), p. 25.
[41] A. Kuo, *The Distance Macro: Preliminary Documentation*, 2nd ed. (SAS Institute, Cary, NC, 1997) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.208).
[42] B. Everitt, *Cluster Analysis* (Heinemann, London, 1974), pp. 9-15.
[43] M. Aldenderfer and R. Blashfield, *Cluster Analysis* (SAGE, Beverly Hills, CA, 1984), p. 38–43.
[44] J. MacQueen, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistical and Probability* (University of California Press, Berkley, CA, 1967) p. 281.
[45] In case of ties (two or more points are at the same minimum distance), one can choose to use the tie-breaking techniques introduced in *SAS/STAT User's Guide, Version 8* (SAS Institute, Cary, NC, 1999).
[46] M. Aldenderfer and R. Blashfield, *Cluster Analysis* (SAGE, Beverly Hills, CA, 1984), p. 36.
[47] R. Springuel, M. Wittmann, and J. Thompson, Applying clustering to statistical analysis of student reasoning about two-dimensional kinematics, Phys. Rev. ST Phys. Educ. Res. **3**, 020107 (2007).
[48] M. Montenegro, G. Aubrecht, and L. Bao, a paper presented at the 2005 Physics Education Research Conference (2005).
[49] F. Baker, *The Basics of Item Response Theory*, 2nd ed. (ERIC, College Park, MD, 2001) (http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/5b/97.pdf).
[50] F. Baker and S. Kim, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed. (Dekker, New York, 2004).
[51] F. Baker, *The Basics of Item Response Theory*, 2nd ed. (ERIC, College Park, MD, 2001), pp. 21–45.
[52] D. Thissen, W. Chen, and D. Bock, MULTILOG-7, Scientific Software International.
[53] R. Bock, Estimating item parameters and latent ability when responses are scored in two or more nominal categories, Psychometrika **37**, 29 (1972).
[54] F. Samejima, University of Tennessee Research Report No. 79-4, 1979.
[55] G. Morris, L. Branum-Martin, N. Harshman, S. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, Testing the test: Item response curves and test quality, Am. J. Phys. **74**, 449 (2006).
[56] Y. Lee, D. Palazzo, R. Warnakulasooriya, and D. Pritchard, Measuring student learning with item response theory, Phys. Rev. ST Phys. Educ. Res. **4**, 010102 (2008).
[57] B. Reeve and P. Fayers, in *Assessing Quality of Life in Clinical Trials: Methods and Practice*, 2nd ed., edited by P. Fayers and R. Hays (Oxford University Press, Oxford, 2005), pp. 55–73.
[58] M. Linacre, Sample size and item calibration stability, Rasch Measurement Transactions **7**, 328 (1994).
[59] L. McLeod, K. Swygert, and D. Thissen, in *Test Scoring*, edited by D. Thissen and H. Wainer (Erlbaum, Mahwah, NJ, 2001), p. 189.
[60] X. Fan, Item response theory and classical test theory: An empirical comparison of their item/person statistics, Educ. Psychol. Meas. **58**, 357 (1998).
[61] D. Harris, Comparison of 1-, 2-, and 3-parameter IRT modes, J. Educ. Meas. **8**, 35 (1989).
[62] L. Bao, Ph.D. thesis, University of Maryland, 1999 (http://www.compadre.org/PER/items/detail.cfm?ID=4760).
[63] L. Bao and E. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).
[64] This is not the only way to deal with missing data. Since there are only two possible scores (1 or 0) a student can get for a particular item, it is reasonable to assign "0" to students for items they do not answer.
[65] A. Agresti and B. Finlay, *Statistical Methods for the Social Sciences*, 4th ed. (Prentice-Hall, Upper Saddle River, NJ, 2009).