Students' proficiency scores within multitrait item response theory

Terry F. Scott* and Daniel Schumayer
Department of Physics, University of Otago, Dunedin 9016, New Zealand
(Received 2 June 2015; published 19 November 2015)

In this paper we present a series of item response models of data collected using the Force Concept Inventory. The Force Concept Inventory (FCI) was designed to poll the Newtonian conception of force viewed as a multidimensional concept, that is, as a complex of distinguishable conceptual dimensions. Several previous studies have developed single-trait item response models of FCI data; however, we feel that multidimensional models are also appropriate given the explicitly multidimensional design of the inventory. The models employed in the research reported here vary in both the number of fitting parameters and the number of underlying latent traits assumed. We calculate several model information statistics to ensure adequate model fit and to determine which of the models provides the optimal balance of information and parsimony. Our analysis indicates that all item response models tested, from the single-trait Rasch model through to a model with ten latent traits, satisfy the standard requirements of fit. However, analysis of model information criteria indicates that the five-trait model is optimal. We note that an earlier factor analysis of the same FCI data also led to a five-factor model. Furthermore the factors in our previous study and the traits identified in the current work match each other well. The optimal five-trait model assigns proficiency scores to all respondents for each of the five traits. We construct a correlation matrix between the proficiencies in each of these traits. This correlation matrix shows strong correlations between some proficiencies, and strong anticorrelations between others. We present an interpretation of this correlation matrix.

DOI: 10.1103/PhysRevSTPER.11.020134 PACS numbers: 01.40.-d, 45.20.D-

I. INTRODUCTION

Assessing knowledge acquisition by students is an unavoidable and necessary part of modern education systems. Multichoice tests are often used for this purpose, especially in large classes. While the primary purpose of these tests is generally to assign a grade, other useful information may be obtained from response data. The value of the information gained by further analyses of test data is of course completely dependent on the quality and purpose of the test. The Force Concept Inventory (FCI) is a preeminent example of such a test and has now been in use in universities and schools for over 20 years [1–3].

A large body of research has been devoted to the analysis of FCI data [4–6], or to the determination of the effectiveness of teaching methods [7–13]. Here we briefly summarize some previous analyses without attempting to be exhaustive.

Wallace and Bailey [12] have shown that IRT models are effective in identifying problematic items in a concept inventory (though not the FCI), such as possible mismatches between item difficulties and student proficiencies, or hidden difficulties with individual items. Planinic *et al.* have also highlighted the importance of examining misfitting items in the Force Concept Inventory [5,14]. Wang and Bao have investigated the applicability and usefulness of a unidimensional, three-parameter item response model for FCI and summarized their findings using a measurement metric. The majority of these findings are confirmed by our single-trait model as well, e.g., we confirm their determinations of the easiest and most difficult items, despite the slight difference between the models analyzed. More importantly Wang and Bao have found that the raw FCI score and the proficiency score are correlated with $R^2 = 0.994$. Finally we mention Han *et al.'s* recent, novel application of IRT analysis [15], whereby the FCI test has been divided into two equal-length tests, covering the same concepts. It has also been suggested that these shorter tests have similar assessment characteristics and produce equivalent test scores to the full FCI test with an overall uncertainty less than 3%. The advantage of splitting the FCI in this way is twofold: first, the test is quicker to administer and second, the split tests may be used in a direct pre- and post-test design without actually reusing the same questions.

This short, and unavoidably very selective, summary demonstrates that educators clearly benefit from investigating new ways to analyze FCI data or from further developing analysis techniques which are already in use.

*Corresponding author.
terry.scott@otago.ac.nz

Our current paper may be considered a companion paper to our earlier analysis of the factor structure of FCI data [13] and here we confirm the results of that paper using alternative techniques. The models we evaluate in this paper are instances of what is often called Item Response Theory (IRT) or Item Response Analysis.

Item Response Theory is a latent variable approach to data analysis and posits that test results are manifestations of underlying "hidden" or "latent" traits possessed by the test respondents. IRT provides techniques for uncovering the properties of these latent variables from test data. Respondents are considered to possess one or more latent traits to a greater or lesser degree and the degree to which they possess these traits determines their response to a particular question. The theory allows for the estimation of characteristic parameters of both respondents and test questions (in standard terminology the latter are called *items*). The characteristic parameter assigned to students is often called *ability* or *proficiency* as this parameter encodes the degree to which each student possesses the relevant latent trait, whatever that trait is.

The number of characteristic parameters relating to the test items depends on the item response model used. In the past, item response models have been restricted to a single latent trait; however, techniques have recently begun to appear which allow for models with multiple underlying traits.

The main purpose of this paper is to complete our analysis of the complex underlying structure which is displayed by FCI response data. As part of this investigation we also present a thorough evaluation of a number of single-trait and multitrait IRT models. We conclude the paper with an analysis of the proficiency scores assigned to students in a multitrait IRT model and we find that this analysis may lead to new insights into the teaching of Newtonian mechanics.

II. DESCRIPTION OF THE DATA

The data used in this study were collected from a physics service course over a two-year period. This is a traditional (i.e., passive student), algebra-based course designed to provide students entering a variety of health science programs (e.g., medicine, dentistry, pharmacy) with the necessary physics for their professional programs. The course consists of six sections: mechanics, bulk materials (i.e., solid and fluid mechanics), thermodynamics, electrostatics, optics, and radiation physics, in this order [16]. In each year more than a thousand students attended this course and sat the test. The students in the course represented a wide spectrum of abilities in physics. The FCI was presented to students via an online course management system at the end of the mechanics section of the course. The students were not required to complete the survey but were unable to finish their internal assessment for the mechanics component of the course until they had at least viewed the FCI. Collecting data for analysis this way does present some issues as there is no control over the time the students take to complete the survey. Moreover, students did not appear to have strong motivation to answer the questions to the best of their ability, since they were not penalized for poor scores in the test. Thus it appeared that there was some risk that the data would be skewed by frivolous attempts. However, such frivolous attempts are easy to filter out. Generally such students merely enter the same response multiple times and then quit the test. Thus we remove from the data set any entry that has the same letter response repeated more than ten times consecutively. The total number of such responses over the two-year period was 20 responses out of a total data set of 2400 students (spoilage $<1\%$). We note that finding all doubtful records would be extremely difficult and we do not claim that we have done so. We have, however, removed the most obviously frivolous attempts at the survey.

III. ITEM RESPONSE MODELS

In this section we describe the main concepts and assumptions of standard item response models, focusing on those statistical methods used in this study. We limit ourselves to a very brief outline of the theory and refer the reader to Refs. [17–21] for more detailed discussions. The section is divided into three parts; the first introduces the concepts and generic assumptions of the item-response approach, the second focuses on the unidimensional formulation, and the third part explains the generalized, multitrait form of IRT.

A. General concepts

Item Response Analysis supposes that a student's response to a particular item in a test depends on an interaction between characteristics of the student and characteristics of the item. Capturing this interaction in mathematical form is the central task of statistical modeling in Item Response Theory. Most models share some properties, e.g., the relevant characteristic of the student is the degree to which that student holds one or more specific underlying properties or traits. In the following discussion we initially consider the simplest case in which there is a single, continuous underlying trait, θ. The degree to which a student possesses this trait quantifies the student's ability to answer a test item correctly. This parameter may take positive or negative values and the calibration of its scale (zero point and interval) is flexible. Often the zero point is chosen to be the sample mean and the scale is fixed by associating $\theta = 1$ with the sample standard deviation.

In these statistical models one would normally assume that the greater the value of θ the higher the probability of a correct response to a question. In a single-trait model we would interpret this trait as being something like "FCI ability." It is better to be noncommittal in the interpretation

of the trait (at least initially), since a label such as "Newtonian-ness" inevitably makes claims about the ability of the FCI to measure the underlying force concept held by students. These presumptions may be reasonable in the sense that a high FCI-ability score may indeed strongly indicate that the student holds a secure Newtonian conception of force. However, it is also quite possible that students with a low FCI-ability score nonetheless have a Newtonian conception of force which is not indicated by their FCI score for extraneous reasons.

B. Unidimensional formulation

The item response theory determines the probability, $P(\theta)$, that student i ($i = 1, 2, \ldots, I$) with a proficiency θ_i chooses the correct answer to the item j ($j = 1, 2, \ldots, J$). A commonly employed expression for this probability is

$$P(x_{ij} = 1|\theta_i, \{a_j, d_j, g_j\}) = g_j + \frac{1 - g_j}{1 + \exp(-a_j\theta_i - d_j)}.$$

This formula includes characteristics of the student, the proficiency θ_i and characteristics of the test item: g_j, a_j, and d_j. The task of item response analysis is to optimally fit these parameters to the observed data. The graph of $P(\theta)$ and the geometric meaning of the parameters are shown in Fig. 1.

As θ_i is a property of the individual student we expect that its value does not change from item to item. In general, we would want students with higher values of θ to possess more developed skills, so that we would expect that they would answer an item correctly with higher probability. This assumption requires that the generic curve is monotonically increasing.

The item characteristic curve shown above is specified by three parameters. These determine the slope of the curve at its midpoint, the position of the midpoint on the abscissa (trait level), and the intercept of the curve with the ordinate (probability) at $\theta \to -\infty$, i.e., when the examinee has no measurable skill for the given trait. In terms of the characteristics of a particular question these parameters may be interpreted as follows. The midpoint of the curve is the point at which the curve reaches half its maximum value. The trait level corresponding to this midpoint is the proficiency at which there is a 50% chance that a correct response will be given to this question. Thus we would expect that 50% of students who have an ability at this trait level would answer this question correctly. This point on the abscissa occurs when $P(\theta) = \frac{1}{2}$ and will depend on all three item fit parameters. However, if $g = 0$ then this point will be given by the value of $-d/a$ and thus this expression is often called the difficulty parameter.

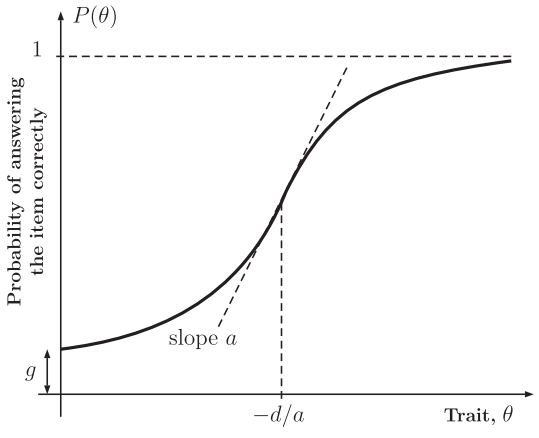


FIG. 1. The standard item response curve of a three-parameter model depicted together with the graphical interpretation of each corresponding parameter, a, d, and g.

The slope of the curve at its midpoint measures the degree to which the question discriminates between students with similar proficiency. If the slope is very large, then two students with similar abilities may have very different probabilities of responding correctly to the question. If the slope is very small then the probability of a correct response does not change greatly between the two students. Thus a large value of the slope parameter indicates a question which discriminates between students who are closely grouped by their proficiency scores. For this reason we call this parameter the *discriminating power* of the item.

Finally, the intercept of the curve with the ordinate gives the probability that a student with very low proficiency nonetheless responds correctly to the question. In other words, this parameter gives the probability that a weak student will guess the correct response to the question, thus g is called *guessing parameter*. While it would be reasonable to expect that g should be as close a possible to the probability of randomly choosing the correct answer (and would thus be determined by the number of options in a multichoice question), it is possible to reduce the guessability of a question with well-chosen distractor options, and it is always possible to increase the guessability with poorly chosen or framed options.

The full three-parameter model is not always necessary and it is often reduced to a one- or two-parameter model by fixing values of some parameter. The one- and two-parameter models are subclasses of the three-parameter model and their functional forms are

$$P^{(1)}(x_{ij} = 1|\theta_i, \{d_j\}) = P(x_{ij} = 1|\theta_i, \{a_0, d_j, 0\}),$$
$$P^{(2)}(x_{ij} = 1|\theta_i, \{a_j, d_j\}) = P(x_{ij} = 1|\theta_i, \{a_j, d_j, 0\}),$$

with the caveat that numerically the parameters of different models are different, since we obtain them by fitting these probability distributions onto the data. Moreover, the value of a_0 is either fixed before the fitting procedure (usually to unity) or treated as another single fitting parameter valid for all items uniformly. The one-parameter model is often called the Rasch model [22], while the

two- and three-parameter models are often called the Birnbaum 2PL and 3PL models [23] in the literature.

The straightforward nature of single-trait item response analysis provides useful estimates of student proficiency which may be used in the more detailed analysis of the latent structures underlying performance in the FCI. Wang and Bao's earlier single-trait analysis [6] of an FCI data set did reveal important information about each item in the test, e.g., relative difficulties. The assumption that there is a single underlying trait is not unassailable and has to be justified by statistical tests in each application of item response analysis. Unidimensional models thus provide an attractively transparent approach for both numerical calculation and the interpretation of an analysis. However, many psychometric tests and several standardized educational tests are multidimensional in that they are constructed from subcomponents that are expected to poll different underlying traits. The FCI was also constructed in this way and is intended to poll six subcomponents of the Newtonian conception of force. These subcomponents are as follows [1]: kinematics, Newton's first, second, and third laws, the superposition principle as applied to forces, and the classification of forces.

In our previous paper [13] we demonstrated that a five-factor model describes our data adequately. It seems, therefore, that a single-trait analysis may leave important structures hidden in the FCI data; thus, we examine models allowing for multiple latent traits.

C. Multitrait formulation

Recently statistical packages for multitrait Item Response Analysis have become available. These models allow for the possibility of multiple underlying traits ($k = 1, 2, \ldots, K$) which influence student response to questions. These underlying traits effectively have the same role as the factors in factor analysis. Multitrait item response theory allows the researcher to perform an exploratory or confirmatory factor analysis, it provides factor weightings as in standard factor analysis, but now the item parameters and student trait levels are also calculated. The functional form of the probability for the multitrait item response model is

$$P(x_{ij} = 1|\boldsymbol{\theta}_i, \{\boldsymbol{a}_j, d_j\}) = g_j + \frac{1 - g_j}{1 + \exp\left(-\boldsymbol{a}_j\boldsymbol{\theta}_i^T - d_j\right)}.$$

The vector $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})$ represents the trait levels (or proficiency levels) of student i, while $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \ldots, a_{jK})$ contains the discriminating parameters for the entire collection of underlying traits for item j. Thus for the kth trait there is both a trait level, θ_{ik}, and a trait discrimination, a_{jk}. Note that there is still a single difficulty parameter, d_j, and guessability parameter, g_j, for each item. In the multitrait model the interpretation of these parameters becomes more complex.

The probability of correctly answering a question now depends on the proficiency of a student in several different traits; these abilities are represented by the trait levels, $\boldsymbol{\theta}_i$, in these traits. In the multitrait item response model, the difficulty parameter, d_j, still encodes the difficulty of the question. Now, however, this difficulty parameter does not simply give the trait level at which a student has a 50% chance of correctly answering a question. Whether or not a student has a 50% chance of answering a question correctly now depends on the sum of the student's trait level in a number of latent traits and thus the 50% point becomes a hypersurface in the multidimensional graph of trait level against probability. A similar interpretation must also be adopted for the guessability parameter, g_j.

The statistical modelling process amounts to fitting such curves to the data and is essentially an optimization procedure. The fitting function is generally a maximum likelihood function, although other methods are also employed [24–26]. A discussion of the various numerical methods used to obtain estimates of the model parameters is beyond the scope of this paper but there are a number of exhaustive texts in the literature which may be consulted for more thorough discussions of technical details [27]. Nevertheless, we note here that the fitting method does require the items to be independent in the sense that the correct solution of one question does not directly depend on the correct solution of any of the other items. The FCI is expected to satisfy this requirement.

Multitrait item response analysis can be cast as a nonlinear extension of classical linear factor analysis methodology, and as such can be employed in an exploratory or confirmatory mode. For exploratory models, the trait axes are constrained to be orthogonal and can be rotated following convergence. If, however, one wishes to check an already existing model no rotational degree of freedom is present, and the IRT analysis becomes similar to confirmatory Factor Analysis. Unidimensional models naturally fall into the confirmatory category as no rotation is possible. In our current study, we did not set the factor structure, therefore, our analysis is entirely exploratory, except in the unidimensional case. In this respect, the current paper is a companion paper to our earlier paper [13] which investigates the factor structure of FCI response data. In our earlier paper we pointed out that there are a number of different techniques, within the Factor Analysis framework, which may be used to identify factors in data. These different mechanisms are often called "rotations." Different rotations may (and very commonly do) assign questions to different factors. Even when questions are assigned to the same factors, different rotations will generally change the factor loadings. This means that the interpretation of these factors must be treated with caution. In our previous paper our interpretation of the factor structure presented was primarily intended to suggest possible avenues for further, more direct, investigation. Multitrait item response

analyses are similar in that they also require an initial stage of Factor Analysis and this analysis is subject to exactly the same constraints. Thus the results of the analysis presented here should again be taken as preliminary, with the primary goal of suggesting fruitful avenues for future research. In particular, the analysis of proficiency scores presented at the end of the paper are intended to suggest the impact of such investigation on the teaching methods employed in this area.

In multitrait Item Response Analysis the factor structure of the data is augmented by a suite of additional parameters. Each question in the survey is assigned a factor loading as is the case in standard Factor Analysis, but now each question is also assigned a difficulty and guessability as well as a discrimination parameter for each underlying factor. Furthermore each student is now assigned a proficiency corresponding to each of the underlying factors. Thus we are now able to analyze the relationship between student understanding of each of the concepts which underlie a full understanding of the Newtonian conception of force. This analysis will be presented at the end of this paper as an example of the power and usefulness of multitrait item response analysis.

A number of commercial and open source packages are available for item response analysis. The analysis presented in this paper was performed using the open source statistical programming language R [28], and the MIRT package [29].

IV. MODEL SELECTION

In this section we discuss the different models used to analyze the data. The most important issue in model selection is the requirement that the model must fit the data. Second, we require some quantitative procedure for comparing different models that all satisfy some minimum fitting criteria. Note that, as with the numerical methods used to implement IRT, model fit and selection is a very large and complex field. We do little more here than indicate which methods were used to identify suitable models; further explanation of the numerical techniques is available in the technical literature; see, e.g., [27].

A. Model fit analysis

Item Response Analysis is performed by fitting a particular model to the data, i.e., finding the values of model parameters which are, in some respect, optimal. It is standard practice to define some "goodness-of-fit" parameter which quantifies the discrepancy between the model and the data, thus a smaller goodness-of-fit value indicates a better fitting model to the observed data.

In this paper we use the root mean square error of approximation (RMSEA) as an industry standard measure of the goodness of fit. The value of RMSEA < 0.05 is employed as an indicator of an acceptable fit and we quote these values for the models discussed in the following. It should be noted that rigorously analyzing the model fit is still a difficult and largely unsolved problem in statistics. Analyzing the behavior of RMSEA statistics in the context of IRT modeling has recently been undertaken by Maydeu-Olivares and a family of generalized statistics has been proposed and examined [30–33]. Although a generic cutoff value for any statistic can be problematic, Maydeu-Olivares and Joe [33] agreed that for RMSEA the 0.05 value can, in most cases, be an acceptable standard.

B. Model comparison

Once it has been established that a group of statistical models all fit the data adequately, it is necessary to choose the most appropriate model for the analysis. This choice may be based on the purpose of the analysis, for example, a single-parameter, single-trait model may be appropriate if all that we are interested in is the relative difficulty of the survey questions. In this paper we will present such a model for exactly this purpose. On the other hand, it may be desired to retrieve as much information as possible from the data and in this case it would appear that the appropriate model would be the one with the most latent traits and fitting parameters. However, as is well known from curve fitting in experimental physics, increasing the number of fitting parameters eventually becomes pathological in that we end up fitting the model to noise rather than capturing the relevant effect. In order to avoid overfitting, a number of statistics has been developed—usually based on some entropic argument—which compares the information content of competing statistical models [34]. These statistics commonly use the likelihood function, L, to compute the information content of the model.

Below we rely on two widely accepted and used measures [35–37]: Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These are defined as

$$\mathrm{AIC} = -2\ln(L) + 2k,$$
$$\mathrm{BIC} = -2\ln(L) + k\ln(n),$$

where k is the number of fit parameters in the model and n is the number of data points with which the model is fit. With these definitions the smaller the AIC or BIC, the better the model. Both AIC and BIC are proportional to the logarithm of the likelihood function, but AIC only penalizes the number of parameters used, while BIC takes into account the size of the data set as well. It is apparent that for larger data sets BIC is the stricter criterion [38]. The best model is then the model which encodes the most information using the least number of fit parameters. In the following section we use the AIC, BIC, and modified versions of both to select the optimal model for our data.

V. RESULTS

In this study we have evaluated a number of statistical models of the data increasing in complexity from a single-trait Rasch model to a ten-trait model. We show that all of these models satisfy a reasonable absolute fit criterion and that the preferred model, from an information theoretical point of view, is the one with five latent traits.

In Table I we list models with increasing number of traits and parameters. All the data are obtained by fixing the relative tolerance of convergence to 10^{-5}, i.e., the optimization procedure is stopped when two consecutive fits resulted in a change in AIC less than this fixed value.

It is clear from Table I that each of the models compared has an adequate absolute fit to the data, in that even the largest value of the RMSEA is less than 0.05. In Fig. 2 we plot several information criteria against the number of traits used in the model. It is clear that the minimum for all of these criteria apart from the AIC occurs with a five-trait model. The AIC has a global minimum at the seven-trait model, but this minimum is extremely flat so that the AIC is not a useful criterion for model selection.

In the rest of this section we will discuss the significance and interpretation of two item response models with a single or with five latent traits.

A. Single-trait model

First we consider a single-trait model. Our treatment will be brief as such models have been discussed by others [5,6,39,40]. This model does, however, provide a benchmark for more sophisticated models. After the concise review of this unidimensional model, we advance to the five-trait model.

TABLE I. Statistical measures of IRT models with varying number of traits are tabulated as calculated by MIRT. The root mean square error of approximation (RMSEA) is provided here to show that all models fit the data acceptably well. The two dominant information criteria, AIC and BIC, are given with the minima underlined. Although the strict minimum AIC value corresponds to $n = 7$, this minimum is quite shallow and may only be the artefact of the optimization procedure. Contrary to AIC, the Bayesian Information Criterion has a well-defined minimum at $n = 5$.

Traits	$\ln(L)$	RMSEA	AIC	BIC
1	−34995	0.0496	70110	70449
2	−34739	0.0421	69655	70158
3	−34556	0.0388	69346	70007
4	−34415	0.0340	69118	69932
5	−34260	0.0262	<u>68860</u>	<u>69821</u>
6	−34230	0.0209	<u>68850</u>	69952
7	−34204	0.0226	<u>68845</u>	70083
8	−34199	0.0200	68882	70250
9	−34206	0.0231	68939	70432
10	−34164	0.0259	68898	70509

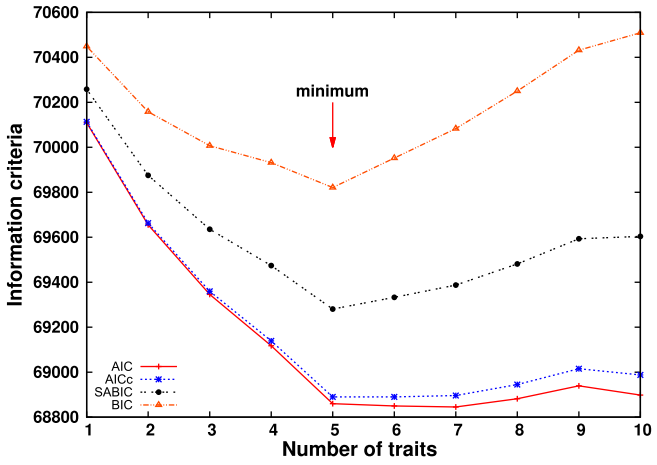


FIG. 2 (color online). Different information criteria are plotted for item response models with varying number of traits. Although Akaike's Information Criterion and its corrected version (AICc) show a broad minimum plateau for $5 \leq n \leq 7$ the other two information criteria have a definite minimum at $n = 5$.

Wang and Bao [6] have fitted a three-parameter model to FCI data collected at the end of a calculus-based introductory mechanics course and have published their estimated item parameters (discrimination, difficulty, and guessing). They found, for example, that the range of difficulty parameters is typically within a range of $(-3, 3)$, and a narrower region of $(-2, 2)$ is covered well by the items. The hardest two items in their model are questions 15 and 26. It is worth mentioning here that differences between the two teaching approaches, their calculus-based course vs our algebra-based course, does not appear to greatly change the structure of the fitted item parameters. Later, Morris *et al.* [40] offered a simplified IRT model called an item response curve method. In this abridged version of IRT the total score of a respondent is used as a proxy for their proficiency level, which tacitly implies that all items are equally weighed. They argued that even their item response curve method is capable of characterizing item difficulty very similarly to that of Wang and Bao's difficulty parameter, b.

Table I indicates that the five-trait model is the preferred model on the basis of a comparison of various information criteria and RMSEA statistics. However, there are often good reasons to accept a simpler model, for example, it may be that all that is required is a quick estimate of student proficiency or an estimate of item difficulty. As shown a single-trait model has an adequate absolute fit to the data since this model has RMSEA < 0.05.

In Fig. 3 we have also depicted the *test* information curve, $I(\theta)$, defined as $I(\theta) = \sum_k I_k(\theta)$, i.e., as the sum of all *item* information curves [19,23]

$$I_k(\theta) = \frac{1}{P_k(\theta)[1 - P_k(\theta)]} \left(\frac{dP_k(\theta)}{d\theta} \right)^2$$

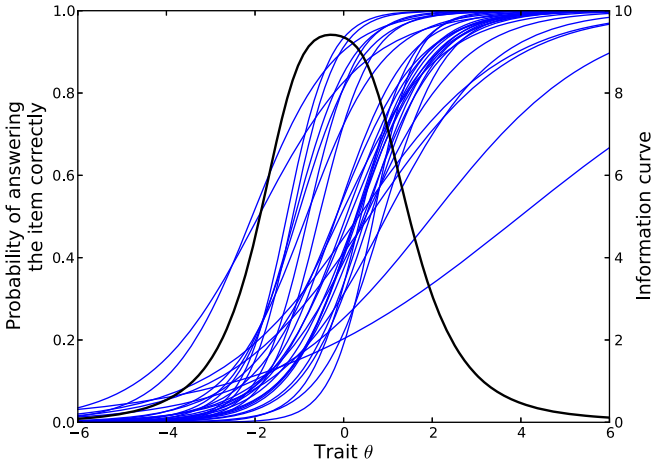


FIG. 3 (color online). The item characteristic curves shown for a single-trait two-parameter model for all 30 items. It is apparent that the $-3 \leq \theta \leq 3$ is well covered by items, and also there are at least two items (question 15 and 26) which poll the students with higher skills. These two items correspond to the two rightmost curves in this graph. The overall test information curve has also been depicted with a heavier line and with its scale denoted on the right vertical axis.

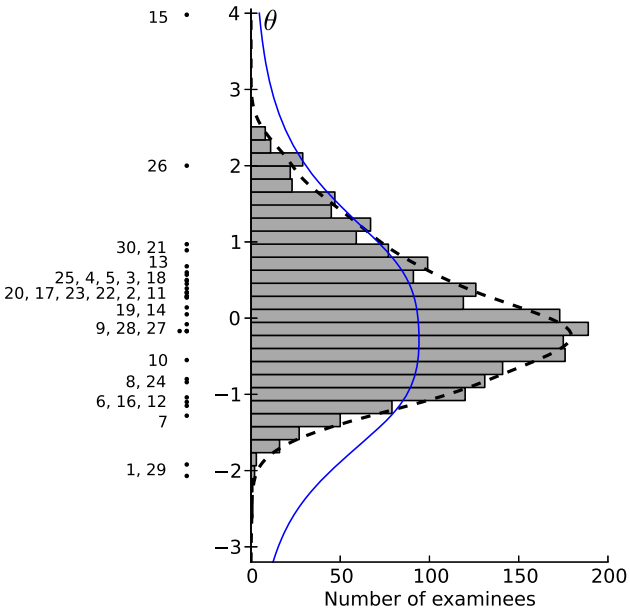


FIG. 4 (color online). The histogram of the students' proficiency scores (right) with the individual item difficulties (left) for the single-trait model. On the right-hand side the smoothed distribution of the skill (bold dashed line) and the test information curve (solid line) have also been depicted.

where the possible parameters of P_k have been dropped from the notation for the sake of transparency. The information function is a measure of the accuracy of the test as a function of trait level. The test information curve is very nearly symmetrical with respect to $\theta = 0$ and its full width at half maximum covers the $-2 \leq \theta \leq 2$ interval, i.e., the FCI seems to poll students equally well with skill below and above average; as can be seen in Fig. 3, there are two item curves (items 15 and 26) falling in the high trait level region and two curves (items 1 and 29) rises in the low trait level region. The test information curve for the FCI is provided here for the first time to the best of our knowledge. Clearly, the FCI is most accurate at moderate trait levels but is less able to accurately assign proficiency scores to very strong or very weak students. This is to be expected and is a result of the relatively tight grouping of item difficulties.

Item Response Theory is a powerful statistical tool in that not only the difficulty of each item, but also the proficiency of the examinees are estimated. The optimization procedure associates as many skill scores as there are latent traits in the model. In the case of a single-factor model, the optimization assigns proficiency score, θ_k to the kth examinee. Since student proficiency and item difficulty are measured on the same scale, it is instructive to compare the distribution of the proficiency scores with the item difficulties themselves.

In Fig. 4 the histogram for the estimated trait scores is plotted on the right-hand side. The solid curve is the same test information curve which is shown in Fig. 3. Here we have rescaled the information curve so it can be compared with the histogram. It is immediately apparent that the test information curve not only covers the skill range of the entire cohort of examinees' but also the test information curve is more or less flat over the bulk of the histogram. This feature is reassuring as it demonstrates that the FCI test accurately determines student proficiency for the bulk of the examinees'.

The item difficulties are shown on the left-hand side of this graph. The same axis is used to indicate the difficulties of the individual items, $-d_j/a_j$ (see Table II). On the left-hand side the item labels have been clustered into sets for the sake of clarity, and within each group the leftmost item has the lowest, while the rightmost item has the highest difficulty.

TABLE II. Fitting parameters for the single-trait model for all 30 items in the FCI test.

j	a_j	d_j	j	a_j	d_j	j	a_j	d_j
1	1.078	2.228	11	1.789	−0.689	21	0.818	−0.793
2	0.613	−0.208	12	1.405	1.461	22	1.241	−0.404
3	0.624	−0.358	13	2.147	−1.453	23	1.290	−0.374
4	0.910	−0.448	14	1.096	−0.148	24	1.848	1.479
5	1.344	−0.677	15	0.344	−1.368	25	1.161	−0.522
6	2.029	2.334	16	1.565	1.724	26	0.542	−1.085
7	2.063	2.650	17	1.147	−0.324	27	0.919	0.072
8	1.142	0.963	18	1.339	−0.802	28	1.272	0.210
9	1.084	0.185	19	1.188	−0.065	29	0.813	1.563
10	1.750	0.966	20	1.345	−0.365	30	1.440	−1.279

B. Five-trait model

The analysis of information criteria, such as the AIC and BIC, suggests that the five-trait model is optimal in that the complexity of the data is captured with the minimum number of fitting parameters. In this section we examine the five-trait model in more detail both in what the model says about the test items and in what it reveals about the respondents. We also compare the structure of these five latent traits to the five factors examined in our previous study [13].

C. Item characteristics

As a first step in analyzing the five-trait model we focus on the item characteristics, i.e., on the $\mathbf{a}_j$ and d_j parameters.

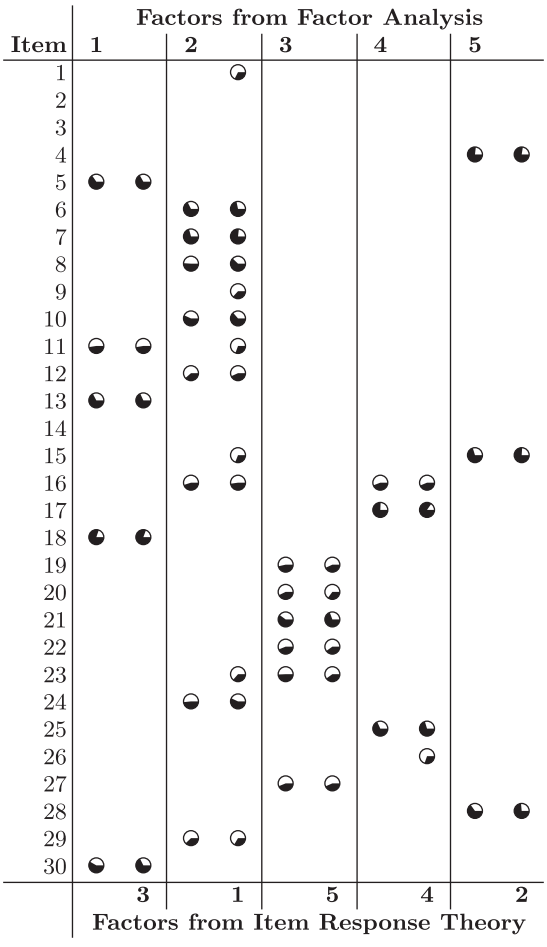
In Table III we give the parameters for a five-trait model. In this model the test items poll student proficiencies in five latent traits. As discussed in Sec. III, each question is assigned a single difficulty parameter, a guessing parameter and five discrimination parameters. An item which does not depend on a particular trait will not load strongly onto that factor and the discrimination parameter associated with that question and trait will be small. As an example let us consider item 2. None of the components of its discriminating parameter, $\mathbf{a}_2 = (-0.453, 0.402, -0.290, 0.061, -0.022)$, are particularly large, and, at the same time, none of the factor loadings of this item are bigger than 0.3, the value chosen to be the cutoff value for appearing in Table IV.

As was noted earlier, the MIRT package performs an exploratory factor analysis to identify the latent traits for the item response analysis.

TABLE III. The fitting parameters, the five discriminating parameters $a_{j1}, \ldots, a_{j5}$, and the difficulty parameter d_j of item j, for all 30 items in the FCI test with the five-trait model. As a reminder, we note here that the individual $a_{j1}, \ldots, a_{j5}$ values cannot directly be compared to the a_j values in Table II (see Sec. III for a detailed explanation).

j	a_{j1}	a_{j2}	a_{j3}	a_{j4}	a_{j5}	d_j
1	−0.842	0.620	0.011	−0.131	−0.241	2.268
2	−0.453	0.402	−0.290	0.061	−0.022	−0.187
3	−0.572	0.297	0.063	−0.278	0.017	−0.348
4	−1.070	0.758	−1.847	−0.027	−0.195	−0.525
5	−1.699	0.169	0.074	0.736	−0.114	−0.760
6	−1.215	1.606	0.495	0.145	−0.752	2.537
7	−1.109	1.842	0.585	0.320	−0.761	2.987
8	−0.723	0.931	0.403	0.281	−0.644	1.046
9	−0.850	0.671	0.335	0.035	−0.187	0.194
10	−1.267	1.167	0.066	0.305	−0.844	1.045
11	−1.685	0.732	0.143	0.497	−0.156	−0.701
12	−0.934	0.916	0.196	0.224	−0.186	1.470
13	−2.780	0.340	0.029	−0.070	0.094	−1.641
14	−0.828	0.743	0.014	−0.187	0.086	−0.123
15	−0.678	−0.201	−1.907	−0.381	0.165	−1.960
16	−0.814	1.606	−0.028	0.633	0.155	1.945
17	−1.103	1.579	0.087	1.174	1.686	−0.484
18	−2.137	−0.12	0.268	0.639	0.044	−0.997
19	−0.908	0.944	0.065	−0.560	0.015	−0.036
20	−1.012	1.002	0.024	−0.347	0.086	−0.344
21	−0.890	0.558	0.181	−1.086	0.625	−0.947
22	−1.200	0.612	0.110	−0.395	0.329	−0.385
23	−1.064	0.915	0.280	−0.500	−0.158	−0.378
24	−1.268	1.205	−0.012	0.054	−0.614	1.538
25	−1.283	0.959	−0.199	0.504	1.520	−0.651
26	−0.422	0.361	0.042	0.026	0.448	−1.113
27	−0.801	0.629	0.254	−0.550	0.000	0.099
28	−1.323	1.455	−1.903	0.000	0.000	0.448
29	−0.350	0.817	0.000	0.000	0.000	1.608
30	−1.937	0.000	0.000	0.000	0.000	−1.414

TABLE IV. The members of all five factors as obtained [13] from factor analysis (top row) and as given independently from item response theory (bottom row). The symbols show which item belongs to which factor with loading larger than 0.3 in magnitude. If no symbol can be found in a row then that item either has not been assigned to a factor or its loading is smaller then 0.3 in absolute value. The black portion within each circular shape is proportional to the loading.



In Table IV we compare the factors obtained for a five-factor item response model with that found in our previous exploratory factor analysis paper (see Table III in Ref. [13]). Each row in the table corresponds to an item in the FCI questionnaire. The circular symbol in a row indicates that the item is part of that factor with loading higher than 0.3 in absolute value. The black shaded area is proportional to the magnitude of loading on that factor, i.e., a full black circle would mean that the item loads entirely on a given factor. Loadings smaller than 0.3 in magnitude are taken to represent weak association and therefore we omit listing such items for that factor.

In order to highlight the structural similarity between the two analyses, the factors from our earlier factor analysis and the current Item Response Theory have been interleaved. The top row in Table IV contains the factor assignment from our previous paper [13] while the last row shows the factor numbering from Item Response Theory. Some resemblance between the results is expected; however, such a striking similarity is surprising. There are some minor differences, e.g., factor analysis would assign item 15 to a single factor (with loading 0.703), item response analysis assigns similar loading to the essentially same factor (loading 0.760), meanwhile it also loads item 15 with a negative loading on to factor 1 (loading −0.310). Such negative loadings suggest that possessing high skills in certain factors may reduce the proficiency on another factor.

Despite such minor differences between the factor assignments, the two statistical approaches result in very similar factors. While this is not necessarily particularly surprising, it is useful in that it allows us to employ the interpretation of these factors which was presented in the previous paper [13] and we will employ the same nomenclature in the present work. For transparency in Table V we list how the factors from the two approaches (factor analysis and item response theory) correspond to each other and how one may describe each factor.

As a final remark we note that items 2, 3, and 14 have not been clearly assigned to any factor in either of the statistical approaches. Items 2 and 14 can be characterized as kinematics questions while 3 is expected to test the understanding of Newton's second law. The fact that these items are left unassigned can be interpreted as these items are superfluous and do not poll the students' skill convincingly. These items could therefore be dropped from the test, we believe, and replaced by three other items.

TABLE V. The factor structure of the FCI data as predicted by two independent statistical models: exploratory factor analysis [13] and item response theory (present work). A description of each factor is also given.

Factor Analysis	Item Response Theory	Description
1	3	Identification of forces
2	1	Newton's first law with zero force
3	5	Newton's second law and kinematics
4	4	Newton's first law with canceling forces
5	2	Newton's third law

D. Student characteristics

One of the attractive features of IRT models is that they are capable of estimating fitting parameters characterizing respondents as well as fitting parameters for test items. Moreover, the item difficulty and the respondents' proficiency scores are measured on the same scale, and are thus directly comparable. In the previous section we analyzed the item characteristics. In this section we will focus on the parameters describing an individual respondent.

The analysis of these proficiency scores is dependent on the particular factor structure employed in the multitrait item response analysis. If a different rotation is employed, the factor structure may well change significantly and thus so will the proficiency scores and the correlations between them. The factor structure we have investigated here is chosen due to the clear interpretation of the meaning of the factors. This clarity leads to an equally clear interpretation of the correlations between the proficiency scores.

Although we have briefly touched upon the students' proficiency scores in an earlier section and compared these scores to the item difficulties, here we would like to examine the students' scores from another direction. As soon as one extends the single-trait model to a multitrait Item Response Theory model, it is inevitable that all examinees are assigned many proficiency scores. Therefore one might also ask whether these proficiency scores show some structure. In other words: Is there any relationship between the trait proficiencies that an examinee exhibits?

Let us hypothesise the following scenario: a cohort is polled on Aristotelian and Newtonian mechanics, and each examinee is assigned two proficiency scores corresponding to these two ideas, θ_A and θ_N, respectively. It is quite plausible that students who have a good understanding of the Aristotelian description of natural motion would struggle with describing the same motion within the Newtonian world view, and vice versa. Therefore one might expect that these two proficiency scores would be anticorrelated with each other.

If there is evidence of correlation structure in the proficiency scores of students taking the FCI, this would indicate relationships between the learning of these concepts. For example, a positive correlation between two trait proficiencies could indicate that an emphasis on one of the concepts in a teaching procedure could also increase understanding of the other concept. Conversely, an anti-correlation could indicate that learning one concept could actively decrease or destabilize the understanding of the other concept. A note of caution is necessary here. As is repeated, correlation is not causation and all we can assert is

TABLE VI. The correlation of students' skill values, $r(\theta_i, \theta_j)$, between the five traits they possess. The traits are identified by indices i and j which take values $i, j = 1, 2, 3, 4, 5$. As the table is symmetric with respect to its main diagonal, only the lower half is provided.

	Traits from Item Response Theory				
	1	2	3	4	5
1	1.000				
2	0.519	1.000			
3	0.769	0.643	1.000		
4	−0.738	−0.602	−0.756	1.000	
5	−0.719	−0.624	−0.770	0.704	1.000

that a correlation of this sort indicates a possible causal link that warrants further investigation.

Based on the five-factor model each student has five skill scores, $\theta_1, \ldots, \theta_5$. We calculated the standard Pearson correlation $r(\theta_i, \theta_j)$ for each pair $(i, j = 1, 2, 3, 4, 5)$. The tabulated result can be found in Table VI.

Table VI shows that there are two groups of trait proficiencies, traits 1, 2, and 3 and traits 4 and 5, with strong positive correlations within groups, i.e., between group members and strong negative correlations between groups, i.e., between members of different groups.

One group contains the proficiencies in factor 1 (Newton's first law with zero force), factor 2 (Newton's third law), and factor 3 (identification of forces). The second group contains proficiencies in factor 4 (Newton's first law with cancelling forces) and factor 5 (Newton's second law and kinematics). We will present an interpretation of this pattern of correlations but we first would like to strongly emphasize that our interpretations of these correlations should be taken as suggestions for further, more direct investigation. We do not claim that our interpretations are directly proven by these correlations.

First we consider the strong positive within group correlations. In the first group, it is not surprising that factors 2 and 3 are combined together since Newton's third law is primarily concerned with the identification of forces, namely action-reaction pairs of forces.

However, it is slightly surprising that factor 1 is also in this proficiency grouping. It would appear that the skill required to answer questions concerning "Newton's first law with zero force" bear some commonality with the skills required to successfully answer questions about Newton's third law and the identification of forces. Furthermore, proficiency in factor 1 is quite strongly anticorrelated with proficiency in factor 4, "Newton's first law with cancelling forces."

Some insight into this issue may be obtained by considering the text of the questions in factor 1. Factor 1 contains 11 questions, one shared with factor 4 (question 16) and another (question 23) shared with factor 5.

Question 16 will not enter into our analysis as it is misplaced in these factors, i.e., it is clearly a question about Newton's third law. It was found to be misplaced in our earlier exploratory factor analysis as well, and in this earlier analysis we propose an explanation. Briefly, we propose in that paper that students analyze this question using a fallacious first law argument which fortuitously provides the correct answer. We will not revisit this discussion here, except to note that the current analysis appears to confirm our earlier factor analysis [13].

Factor 1 contains questions which appear to favor visual problem solving modalities. In factor 1, questions 6, 7, 8, 12, and 23 are "visual choice" questions in that they all require the student to choose the correct diagram from a set. There are only two other questions of this type in the FCI, questions 14 and 21. It is striking that nearly half the questions in factor 1 are of this type, and this type of question appears almost exclusively in factor 1. We would therefore suggest that there is a strong visual element in this factor in that these questions could be correctly answered by a student with a strong visual memory and some experience in situations similar to those depicted in these questions. The fact that proficiency in this factor is strongly correlated with proficiency in factors 2 and 3 suggests that research into the use of visual thinking and problem solving in the identification of forces and the understanding of Newton's third law would perhaps be fruitful.

We note that the other questions in factor 1 tend to be "follow up" questions to these visual choice questions. For example, questions 9, 10, and 11 ask further questions about the diagrams presented in question 8. These questions do appear to be independent in the sense required for item response analysis; the answer to question 8 is not required to answer questions 9, 10, and 11. However the diagrams in question 8 do provide a significant "visual aid" to a student attempting questions 9, 10, and 11. A student with a strong visual memory who correctly chooses option B as the answer to question 8 is then able to use that diagram to assist them in answering the later questions. Further understanding of Newton's first law is required, but option B does supply further cognitive input in the reasoning process.

Now we consider the strong positive correlation between proficiencies in factors 4 and 5. It is not surprising that the proficiency in these two factors is strongly correlated since Newton's first law is a special case of Newton's second law. The surprise is not that proficiencies in factors 4 and 5 are in this group; the surprise is that proficiency in factor 1 is not also in this group. As discussed above, it appears that there may be a strong visual element in proficiency in factor 1 questions. The grouping of this proficiency with factors 2 and 3 rather than with factors 4 and 5 suggests that the visual character of this proficiency is more closely correlated with the skills required to solve factor 2 and 3 problems than with the clear similarity between the physics of factor 1 and the physics of factor 4. The correlation would appear to be due to similarity in thought process rather than the similarity in content.

Finally we will discuss the strong anticorrelation between the two groups of factor proficiencies, that is, between the group containing factors 1, 2, and 3 and the group containing factors 4 and 5.

To begin with we should point out that the main result is that there is an anticorrelation between these two groups. Further research is required to determine the cause of this anticorrelation. We suggest possible reasons for this anticorrelation with this caveat in mind. We suggest two possible sources of the observed anticorrelation; these two suggestions are not mutually exclusive.

As we have already discussed, one possibility is that the factors are successfully solved by different styles of thinking. Proficiency in factor 1 seems to involve visual problem solving, whereas factors 4 and 5 do not so clearly depend on this diagrammatic mode of thought. If this interpretation is correct, then the separation between the two groupings may be due to differences between the required problem solving modes and conflicts between the operation of these two modes. This interpretation would also highlight the importance of developing a skill for switching between multiple representations of a physics problem [41]. Our hypothesis regarding the visual difference between the latent FCI factors is partially supported by a previous study [42], which found positive correlation between students' preinstruction level of representational consistency and their learning of forces.

A second possibility is that the physics involved in the two groups conflicts. This would mean that understanding Newton's third law destabilizes understanding of Newton's second law. Learning new concepts challenges old knowledge. This is an unavoidable character of learning. Thus it would not be surprising if learning Newton's third law decreased a student's understanding of Newton's second law, at least, temporarily.

VI. CONCLUSION

Concept inventories are frequently used to assess students' proficiencies in a particular subject and to provide feedback about teaching practices. The analysis presented in this paper highlights the value of item response models in physics education research. The development of multitrait item response models in particular allows for the extension of factor analysis techniques to include the assignment of proficiency scores to respondents. This in turn allows for the investigation of the multiple skills students employ to complete surveys like the FCI.

In this paper we have presented a detailed analysis of several single and multitrait item response models of FCI data. It has been shown that a simple single-trait, single-parameter model may be applied to the data with adequate fit. It has further been shown, via the use of several standard information criteria, that the optimal item response model of the data has five underlying traits and seven item parameters. These five traits have been shown to correspond closely to the factors found in an exploratory factor analysis presented in our previous paper [13].

Finally, we have analyzed the proficiency scores assigned to respondents for each trait in the five-trait model. We then constructed a correlation matrix between the trait proficiencies in these five traits. This matrix showed that there are two groups of trait proficiencies. Within each group the trait proficiencies are quite strongly and positively correlated with each other. Between groups the trait proficiencies are quite strong and negatively correlated with each other. We have suggested that the strong correlation within one group suggests a significant visual component to these trait proficiencies, and since one of the trait proficiencies in this group is Newton's third law trait proficiency, we suggest that a strong proficiency in this concept relies on good visual problem solving. The strong and negative correlation between the two groups is interpreted as being attributable to differences in problem solving modalities, and also the possibility that an understanding of Newton's third law challenges a preexisting understanding of Newton's second law. We point out that these interpretations should be taken primarily as suggestions for future research.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Department of Physics, University of Otago, and by the MBIE UOOX-1208.

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

[2] Authorized educators can request the latest FCI test from modeling.asu.edu/R&E/Research.html.

[3] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).

[4] 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).

[5] 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).

[6] J. Wang and L. Bao, Analyzing force concept inventory with item response theory, Am. J. Phys. **78**, 1064 (2010).

[7] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting force concept inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).

[8] A. Savinainen and J. Viiri, The force concept inventory as a measure of students conceptual coherence, Int. J. Sci. Math. Educ. **6**, 719 (2008).

[9] 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).

[10] 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).

[11] P. Nieminen, A. Savinainen, and J. Viiri, Erratum: Force Concept Inventory-based multiple-choice test for investigating students' representational consistency, Phys. Rev. ST Phys. Educ. Res. **6**, 029903 (2010).

[12] C. S. Wallace and J. M. Bailey, Do concept inventories actually measure anything?, Astron. Educ. Rev. **9**, 010116 (2010).

[13] 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).

[14] M. Planinic, W. J. Boone, R. Krsnik, and M. L. Beilfuss, Exploring alternative conceptions from Newtonian dynamics and simple DC circuits: Links between item difficulty and item confidence, J. Res. Sci. Teach. **43**, 150 (2006).

[15] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[16] K. Franklin, P. Muir, T. Scott, L. Wilcocks, and P. Yates, *Introduction to Biological Physics for the Health and Life Sciences*, 1st ed. (Wiley, New York, 2010).

[17] F. M. Lord, Practical applications of item characteristic curve Theory, J. Educ. Measure. **14**, 117 (1977).

[18] F. M. Lord, Applications of Item Response Theory to Practical Testing Problems (Lawrence Erlbaum Associates, Inc, Hillsdale, 1980).

[19] R. K. Hambleton, *Fundamentals of Item Response Theory*, Measurement Methods for the Social Science (SAGE Publications, Newbury Park, Calif, 1991), Vol. 2.

[20] P. R. J. de Ayala, *The Theory and Practice of Item Response Theory*, Methodology in the Social Sciences (The Guilford Press, New York, 2008).

[21] C. DeMars, *Item Response Theory, Understanding Statistics: Measurement* (Oxford University Press, Oxford, 2010).

[22] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests* (University of Chicago Press, Chicago, 1980).

[23] A. Birnbaum, *Statistical Theories of Mental Test Scores* (Addison-Wesley Publishing Company, New York, 1968). Some latent trait models and their use in inferring an examinee's ability.

[24] R. J. Patz and B. W. Junker, A straightforward approach to Markov Chain Monte Carlo methods for item response models, J. Educ. Behav. Stat. **24**, 146 (1999).

[25] T. R. Johnson, Item response modeling with sum scores, Appl. Psychol. Meas. **37**, 638 (2013).

[26] S. Monroe and L. Cai, Estimation of a Ramsay-Curve item response theory model by the Metropolis–Hastings Robbins–Monro Algorithm, Educ. Psychol. Meas. **74**, 343 (2014).

[27] F. B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Techniques*, 2nd ed., Statistics: A Series of Textbooks and Monographs (Marcel Dekker, New York, 2004).

[28] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2014).

[29] R. P. Chalmers, mirt: A multidimensional item response theory package for the R environment, J. Stat. Software **48** (2012).

[30] A. Maydeu-Olivares, Goodness-of-Fit assessment of item response theory models, Meas. Interdiscip. Res. Perspect. **11**, 71 (2013).

[31] D. Thissen, The meaning of Goodness-of-Fit Tests: Commentary on "Goodness-of-Fit assessment of item response theory models", Meas. Interdiscip. Res. Perspect. **11**, 123 (2013).

[32] L. Cai and S. Monroe, IRT model fit evaluation from theory to practice: Progress and some unanswered questions, Meas. Interdiscip. Res. Perspect. **11**, 102 (2013).

[33] A. Maydeu-Olivares and H. Joe, Assessing approximate fit in categorical data analysis, Multivariate Behav. Res. **49**, 305 (2014).

[34] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 2008).

[35] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control **19**, 716 (1974).

[36] H. Akaike, in *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, edited by E. Parzen, K. Tanabe, and G. Kitagawa (Springer, New York, 1998) pp. 199–213.

[37] G. Schwarz, Estimating the dimension of a model, Ann. Stat. **6**, 461 (1978).

[38] Note that some texts define AIC with the opposite sign, i.e. $\mathrm{AIC} = -2k + 2\ln(L)$; therefore, one should choose the model with the largest value of the AIC statistic.

[39] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. 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).

[40] 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).

[41] D. E. Meltzer, Relation between students' problem-solving performance and representational format, Am. J. Phys. **73**, 463 (2005).

[42] 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).