Dividing the Force Concept Inventory into two equivalent half-length tests

Jing Han,[1] Lei Bao,[1,2,*] Li Chen,[1,3] Tianfang Cai,[2] Yuan Pi,[4]
Shaona Zhou,[5] Yan Tu,[3,†] and Kathleen Koenig[6,‡]

[1]*The Ohio State University, Columbus, Ohio 43210, USA*
[2]*Beijing Jiaotong University, Beijing 100044, China*
[3]*Southeast University, Nanjing, Jiangsu 210096, China*
[4]*Central China Normal University, Wuhan, Hubei 430079, China*
[5]*South China Normal University, Guangzhou, Guangdong 510631, China*
[6]*University of Cincinnati, Cincinnati, Ohio 45220, USA*

(Received 22 July 2014; published 4 May 2015)

The Force Concept Inventory (FCI) is a 30-question multiple-choice assessment that has been a building block for much of the physics education research done today. In practice, there are often concerns regarding the length of the test and possible test-retest effects. Since many studies in the literature use the mean score of the FCI as the primary variable, it would be useful then to have different shorter tests that can produce FCI-equivalent scores while providing the benefits of being quicker to administer and overcoming the test-retest effects. In this study, we divide the 1995 version of the FCI into two half-length tests; each contains a different subset of the original FCI questions. The two new tests are shorter, still cover the same set of concepts, and produce mean scores equivalent to those of the FCI. Using a large quantitative data set collected at a large midwestern university, we statistically compare the assessment features of the two half-length tests and the full-length FCI. The results show that the mean error of equivalent scores between any two of the three tests is within 3%. Scores from all tests are well correlated. Based on the analysis, it appears that the two half-length tests can be a viable option for score based assessment that need to administer tests quickly or need to measure short-term gains where using identical pre- and post-test questions is a concern.

DOI: 10.1103/PhysRevSTPER.11.010112 PACS numbers: 01.40.Fk, 01.40.gf

I. INTRODUCTION

Assessing what students learn or what they know is an important but difficult task in education research. In the physics education community, the Force Concept Inventory (FCI) is the most often used tool of assessment [1,2]. There are two versions of the FCI, the original version published in 1992 and the revised version released in 1995. In this study, the 1995 version is used, which contains 30 multiple-choice questions covering topics commonly taught in introductory mechanics. These questions were designed to probe conceptions that are shown to be common among high school and college students [3,4,5].

In education research and evaluation, pre-post testing is a very popular assessment method. For example, in Hake's study, over 6000 students were pre-post tested using the FCI [2]. The results showed that interactive engagement classroom settings (ones that included hands-on learning and discussions) enhanced students' learning as opposed to traditional classroom settings characterized by lectures. These are powerful results for educators and have influenced much of the physics education research and curriculum development that takes place today.

When doing assessment in real classroom settings, instructors and researchers are often concerned by a number of practical issues. The first is the time taken for students to complete the assessment. The typical 40-minute allotted time for the FCI takes up nearly an entire class period, which may affect the willingness of instructors to use it for pre-post testing in their already crowded schedules. However, less time is often not an option since test anxiety could be raised by any reduction of the time allotted [6]. As a comparison, the Chemistry Concept Inventory was developed with only 20 questions so that it could be administered in a short period of time [7]. The second issue is the test-retest memorization effect when identical questions are used in pre-post testing. It has been shown that this test-retest memory effect tends to fade over a time period longer than 5 weeks [8–10]. If researchers want to study short-term learning gains, the test-retest issue needs to be carefully addressed [11]. A possible solution to measuring short-term learning gains is to use equivalent parallel tests for pre-post testing.

Therefore, it is advantageous to have shorter parallel tests that have similar assessment capacities and are quick to administer. Since the FCI is the most widely used

*bao.15@osu.edu
†tuyan@seu.edu.cn
‡koenigkn@ucmail.uc.edu

research tool in PER, with which researchers have produced a large collection of data and research outcomes, it would be ideal to create tests equivalent to the FCI so that the results from the new tests can be compared with the existing work.

The FCI test measures a range of concepts in force and motion and is designed to have multiple questions for each concept. In this study, we use both content analysis of the FCI questions and descriptive statistics of the existing data to split the FCI test into two equivalent short version tests (nearly half-length) that each contains 14 questions. The reduction in numbers of questions will adversely impact the assessment capability in measuring individual concepts. Therefore, the focus of this study is aimed at producing equivalent and scalable total scores of the different versions of tests.

A timed student trial showed that the majority of students were able to complete the short tests in less than half the time needed for the full FCI. We then conducted extensive quantitative analysis to compare the assessment characteristics of the three versions of the tests (two half-length tests and the full-length FCI test) and to determine the typical measurement uncertainties. The assessment parameters obtained in this study establish an important baseline for using the short version tests in practice.

II. METHOD AND DESIGN

As an overview, the method in this study includes five general steps. First, a team of physics education researchers hand picked the FCI questions into two short tests aimed to measure approximately the same set of concepts and produce mean scores that can be equated with the full-length FCI scores. Second, using a large scale FCI data set, a computational regression process is used to sort through the possible combinations of the two short tests to identify the optimal construction that minimizes the total errors among the mean scores of the different tests. Third, the Item Response Theory (IRT) is used to estimate and compare the assessment features of the two short tests and the FCI test, which suggest that simple linear models can be used to convert the mean score of one test to an equivalent mean score of another test. In the fourth step, actual linear conversion models are determined to convert scores of one test to another. The uncertainties in such conversions are also evaluated. As the final step, which evaluates the reliability of the method, the conversion models are applied with data sets from very different populations and the overall uncertainties in such applications are evaluated to obtain an approximate scale of error tolerance for using the new tests in practice.

In the first step of this study, the FCI questions were grouped into 7 clusters (see Table I) by a team of content experts in the field of physics education research including 5 professors and 6 graduate students. The grouping took into account a wide range of factors concerning the design of the FCI [1,3–5,12,13], the involved concepts and contexts of the questions [13–15], as well as the construction features of the test such as individual and sequenced question structures.

TABLE I. The physics concepts assessed in the two half-length tests, HFCI1 and HFCI2. The numbers listed under each test reflect question numbers on the full version of the FCI.

Concept	HFCI1	HFCI2
Free fall	2	2
Newton's third law	15, 28	4, 28
Force motion	13, 17, 26	25, 26, 30
Circular motion	5, 6	7, 18
Projectile motion	12	14
Kinematics	19	20
Force motion cluster	8–11	21–24

The expert team then hand-picked questions in each cluster into two groups to form the initial versions of the two half-length tests, referred to as HFCI1 and HFCI2. The splitting of questions was determined by the expert team based on a number of practical considerations regarding contexts, average scores, question sequences, etc. This splitting process produced the initial candidate versions of the two half-length tests as well as the constraints for certain questions to go into either of the tests. For example, only questions in the same content clusters can be switched between the two tests. In addition, questions in a sequence needed to remain in the same sequence. A computer script was written to scan all allowed question swaps as well as removing up to a total of two FCI questions between HFCI1 and HFCI2 to find the most optimized structures of the two tests so that the sum of differences between scores of the two tests in both pre- and post-test applications was minimized. The final results of splitting of the questions through computer and manual selections are listed in Table I.

As shown in Table I, three FCI questions (2, 26, and 28) are used on both half-length tests for test equating purposes. These three questions cover simple projectile (free fall) motion, force and motion, and Newton's third law, allowing a common base for test equating calculations. Questions 1, 3, 16, 27, and 29 are not used on the new tests in part due to the limited length of the two half-length tests. In addition, questions 1 and 3 often have the highest scores on the FCI test resulting in the ceiling effect on the pretest for college students and on the post-test for high school students. Questions 15 and 16 are two sequenced questions on the concept of Newton's third law using identical contexts. The results of question 16 can be significantly influenced by question 15 [13]. Therefore, question 16 is not used in the half-length tests.

Question 27 is the last question in a 3-question sequence on the concept of force and motion using the context of box pushing. Performance on this question is significantly

higher than the first two questions in the sequence. Based on the computation results described in the second step above, assigning question 27 to either of the half-length tests would significantly impact the measurement equivalence of the two tests. As a result, question 27 is not used in the half-length tests. Question 29 involves the types of forces on an object and is the only question in this category. There hasn't been much established research on student difficulties in this specific area. In addition, quantitative analyses showed that including or removing this question doesn't change the assessment characteristics of the tests. Therefore, to limit the lengths and overlaps of the two half-length tests, question 29 is also removed.

Although from the expert's point of view, the two half-length tests span the same sets of physics concepts, the two groups of questions involve different contexts. Existing research on context sensitivity of assessment questions has shown that contextual features of questions can significantly influence how students respond to questions [14,15]. This leads to concerns on the reliability and equivalence of the new tests. In this study, we will evaluate the equivalence and reliability of the mean scores of the new tests using results from a large-scale statistical analysis. The results of this evaluation only apply to mean scores of the tests and are not intended to be used to make implications on the equivalence regarding student understanding of specific conceptual domains.

Ideally, equivalent instruments not only produce equivalent scores but also have similar assessment characteristics such as discrimination, difficulty, and guessing chances. In this research, the statistical analysis focuses on the equivalence and reliability of the three tests (two half-length tests and the original FCI). First, we use Item Response Theory to estimate and compare the basic assessment features including the discrimination, difficulty, and students' guessing parameters of the three tests. We then quantitatively determine the numerical models for score conversions among the different tests. If the differences among the assessment features and the errors produced in the score conversions are both acceptable, the three tests are then considered to be statistically equivalent and can be used interchangeably in practice.

The data used in this analysis were collected at a large midwestern state university for a period of 5 years, the students who enrolled in calculus-based introductory mechanics courses took the full FCI as a pretest during the second week and as a post-test in the week before final exams, a time elapse of approximately nine weeks. The average pre- and post-test scores for each quarter remained fairly steady over time. Therefore, we treated the students from different years as the same student population. The data set consists of unmatched pre- and post-test scores, containing 3139 pretest scores and 2526 post-test scores. A few matched data sets from the same population have also been experimented. The differences between matched and unmatched data are small (less than 1% among all tests).

The data include results for all 30 questions of the FCI. Based on the question assignments listed in Table I, data on selected questions were used to calculate the results on HFCI1 and HFCI2. In Table II, the basic statistics of all

TABLE II. Basic descriptive statistical comparisons of the two half-length tests and the original FCI test. All scores and standard deviations are in the scale of a 100-point score. The score changes (ΔS) are calculated based on the average scores of the pre- and post-test.

	Pretest			Post-test			Pre-post score change (ΔS) and normalized gain (g)		
	HFCI1	HFCI2	FCI	HFCI1	HFCI2	FCI	HFCI1	HFCI2	FCI
Concept areas	Score	Score	Score	Score	Score	Score	ΔS (g)	ΔS (g)	ΔS (g)
Free fall	43.77	43.77	66.43	62.27	62.27	77.42	18.5 (32.90)	18.5 (32.90)	10.99 (32.74)
Newton's third law	33.50	33.83	41.00	56.12	64.75	64.87	22.62 (34.01)	30.92 (46.72)	23.87 (40.46)
Force motion	18.06	17.69	27.80	44.22	41.90	50.88	26.16 (31.92)	24.21 (29.41)	23.09 (31.97)
Circular motion	50.32	46.11	48.22	67.52	71.34	69.43	17.2 (34.62)	25.22 (46.81)	21.21 (40.96)
Projectile motion	78.31	61.07	69.69	88.24	69.04	78.64	9.94 (45.80)	7.97 (20.48)	8.95 (29.53)
Kinematics	67.44	65.50	66.47	70.51	72.96	71.73	3.06 (9.41)	7.46 (21.63)	5.26 (15.69)
Force motion cluster	50.29	56.98	53.64	67.64	66.22	66.93	17.35 (34.90)	9.24 (21.47)	13.29 (28.67)
Test total (SD)	43.75 (19.79)	43.66 (21.15)	49.26 (18.17)	62.25 (21.55)	61.93 (23.18)	66.23 (19.22)	18.5 (32.89)	18.27 (32.43)	16.97 (33.45)

three tests are summarized, including the average pre- and post-test scores, pre-post score changes, and pre-post normalized gains of different concept areas and the whole tests. The results show that the total scores of the two half-length tests are virtually identical; the difference is 0.09% on pretest and 0.32% on post-test. Because of the removal of several high-scoring FCI questions, the scores of the half-length tests are typically 5% lower than that of the FCI. Meanwhile, larger variations are observed among the scores of different tests in some of the individual concept areas. This can be caused by at least two factors: the small number of questions in a single concept area and the sensitivity of student performance to question contexts. As a result, it is recommended that only the total scores and score changes are used for evaluating student performances.

Between pre- and post-test, the raw score changes for the two short tests are also very similar, with a difference of 0.23%. Between the short tests and the FCI, the raw score changes differ on the order of 1.5%. Similar results are also observed with the normalized gains. These results suggest that on the basis of descriptive statistics, the total scores, score changes, and normalized gains of the two short tests provide equivalent measures of student performances and learning gains. The FCI scores are slightly but consistently higher than those of the short tests. In the later part of this paper, a numerical relation will be quantitatively determined, which can convert the score of one test to an equivalent score of another test. With such conversions, the two short tests and the FCI can be used interchangeably to measure student overall performances.

III. TEST EVALUATION USING IRT ANALYSIS

When evaluating the assessment equivalence of different tests, the mean score is often less important than the discrimination and measurement scale. For example, if two tests produce different mean scores with the same population but have similar discriminations and measurement ranges, the test scores can be easily equated with a constant offset. In this study, a standard approach of the IRT-based test equating analysis is used to evaluate the score-based assessment equivalence of the half-length tests and the full-length FCI test.

Test equating is a well-established field that uses multiple models and methods to study the reliability and equivalence of assessment instruments [16,17]. IRT is often used as the theoretical basis for advanced test equating analysis that can produce assessment parameters based on a probabilistic framework. The typical set of assessment parameters includes test discrimination, test difficulty, and the guessing factor. Existing research has shown that IRT can be used to evaluate features of the FCI test [18,19]. The advantages of using an IRT-based method is that it can maintain assessment consistency when student populations have very different mean scores and that the estimated assessment parameters can help extend the assessment scale into questions not used on the test [16]. In this section, we apply the three-parameter IRT model to estimate and compare the assessment parameters of the two half-length tests and the full-length FCI test. The model is given by

$$P(\theta) = c + \frac{1-c}{1+\exp[-1.7a(\theta-b)]}. \quad (1)$$

Here, $P(\theta)$ is the probability of a student with ability θ to correctly answer a question, which is the model predicted score of the student. The assessment characteristics of a question are described in terms of three parameters: the item discrimination a, the item difficulty b, and the guessing parameter c. All the parameters in Eq. (1) are between 0 and 1 and are estimated with a large-scale data set through a regression process using marginal maximum likelihood (MML) algorithms that minimize the errors between observed student scores and model predicted scores.

To apply the IRT model, the data should follow a normal distribution [20]. Analysis of the data set used in this paper suggests that student scores on the two half-length tests reasonably follow a normal distribution ($R^2 > 0.97$) and that the condition of this data set is appropriate for conducting IRT analysis. Similar results have also been reported in our previous study on IRT application to FCI data [18].

Using the three-parameter IRT model, the assessment parameters of the three tests were calculated with pretest data. The use of pretest data is based on the implementation model that treats the test instrument as a fixed entity and the students with changing abilities between pre- and post-test. The IRT model estimates the three parameters for each item on a test. The results were then averaged to produce the parameters of the test listed in Table III [21]. Overall, the assessment features of the three tests are similar in terms of discrimination and guessing. However, the difficulty level of the full FCI is smaller than that of the two half-length tests, which is consistent with the results of test scores shown in Table II.

For all three tests, the guessing chances are on the order of 10%–15%, which is below the structural uncertainty of the five-choice single-response questions. This result is consistent with the existing literature, which shows that the distractors of the FCI tests are not equally attended to by students due to students' naïve understanding of the related

TABLE III. Assessment parameters for the two half-length tests and the full-length FCI calculated using IRT model applied on the pretest score.

	Discrimination a (SD)	Difficulty b (SD)	Guessing chance c (SD)
HFCI1	1.86 (0.95)	0.59 (1.08)	0.15 (0.14)
HFCI2	1.74 (0.92)	0.51 (0.91)	0.11 (0.09)
FCI	1.65 (0.95)	0.26 (1.09)	0.13 (0.12)

physics concepts [14]. As a result, students holding certain naïve conceptions would consistently chose a specific subset of incorrect answers reducing the occurrences of random guessing.

The discrimination parameters of the two new tests are similar and are slightly higher than that of the FCI. The differences are about 0.2, which would not significantly alter the response characteristics of the tests. To explore how variations of the three assessment parameters might impact the response probabilities, the response curves of the FCI and the two half-length tests were plotted in Fig. 1 along with an adjusted curve of HFCI1, which will be discussed later. In Fig. 1, the solid black line represents the response curve of the FCI. The two nearly overlapping dotted and dashed black lines are the response curves of HFCI1 and HFCI2. This shows that the two half-length tests have nearly identical response characteristics concerning their difficulty levels and discriminations. The chances of guessing the HFCI1 are slightly larger than that of the HFCI2 at low θ, which in this case will not impact the overall assessment properties as the ability measure θ of most students is in the range from -1 to $+2$.

Comparing the HFCI tests and the FCI, their two response curves look very similar except that they are shifted horizontally by a constant equal to the difference between their difficulty parameters (=0.33 for the HFCI1, as an example). The HFCI tests are slightly harder and therefore their response curves are shifted towards the high-θ direction.

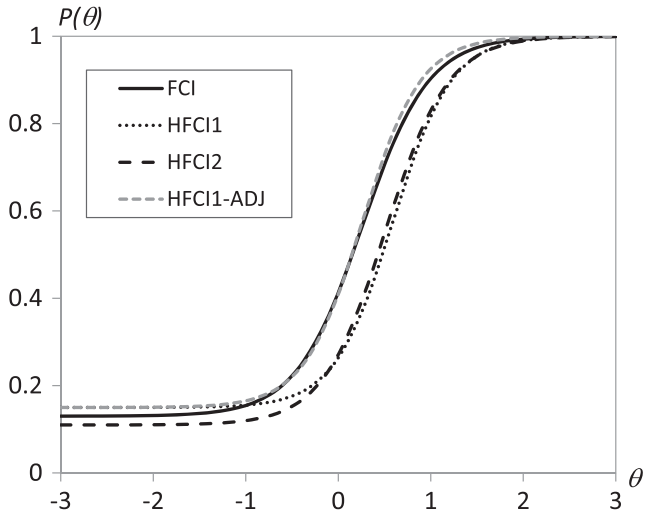

FIG. 1. IRT response curves of FCI and HFCI1. The horizontal axis represents the scale of the student ability parameter θ, which is typically between -3 and $+3$. The majority of students fall in the range of -1 to $+2$. The vertical dimension gives the probabilities of correct response (or the mean score) on the test of students with specific values of ability θ. The solid line plots the response curve of the FCI test based on the parameters in Table III. The black dotted and dashed lines plot the response curve of HFCI1 and HFCI2. The gray dashed line plots the response curve of HFCI1 with its difficulty parameter reduced by 0.33.

Using HFCI1 as an example to compare with the FCI, the features of the two response curves are ideal for test equating using linear models. Since the main difference between the two curves is a constant shift, one can quantitatively adjust the difference of the difficulty parameter to make good predictions from the score of one test to the other. As an example, an adjusted HFCI1 response curve is produced by subtracting the difficulty parameter of HFCI1 with 0.33. The curve is plotted with a gray dashed line in Fig. 1. The results show a nearly identical curve to the FCI's response curve. The standard deviation of the differences in predicted student scores produced by the response curves of the FCI and the adjusted HFCI1 is 1.4%, which provides a rough estimation of the approximate uncertainty of score equating between FCI and HFCI1.

In summary, the results in Table III and Fig. 1 suggest that by using linear models of test equating the three tests can produce equivalent assessment scores with errors at the level of 1.4%. Therefore, the two new tests can be alternative score based assessment options for replacing the FCI.

IV. TEST EQUATING RESULTS

Based on the descriptive statistics and IRT analysis, the two half-length tests and the original FCI have similar assessment features and can produce equitable measurement outcomes. In practice, it is often helpful to convert the scores of the new tests to equivalent FCI scores, which allow direct comparisons of the new results with the large amount of FCI data in the literature. In this section, the conversion scales are determined so that the score measured by any one of the three tests can be converted into an equivalent score on another test.

It is worth clarifying that score conversions operate on the average total scores of the different tests and don't apply to individual student scores due to the multifaceted uncertainties within the individual students' problem-solving processes. The average of total test scores of a sizable sample ($N \sim 100$) can often reduce the impact of the individual uncertainties and produce results that are statistically valid and reliable.

With three versions of tests and two testing conditions (pre and post), a total of six scatter plots are produced in Fig. 2 to show the relations between the scores of any two tests in different testing conditions. A point in the figure gives the average scores on two tests for a group of students in a particular bin. The bins are defined with the scores of the test labeled on the horizontal axis of each curve. Since each half-length test has 14 questions, there are a total of 15 score bins. Students are assigned to one of the 15 bins based on their scores on the test indicated on the horizontal axis. For the group of students in a particular bin, their average scores on the two tests indicated on the horizontal and vertical axes are calculated and used to position the point on the figure. The

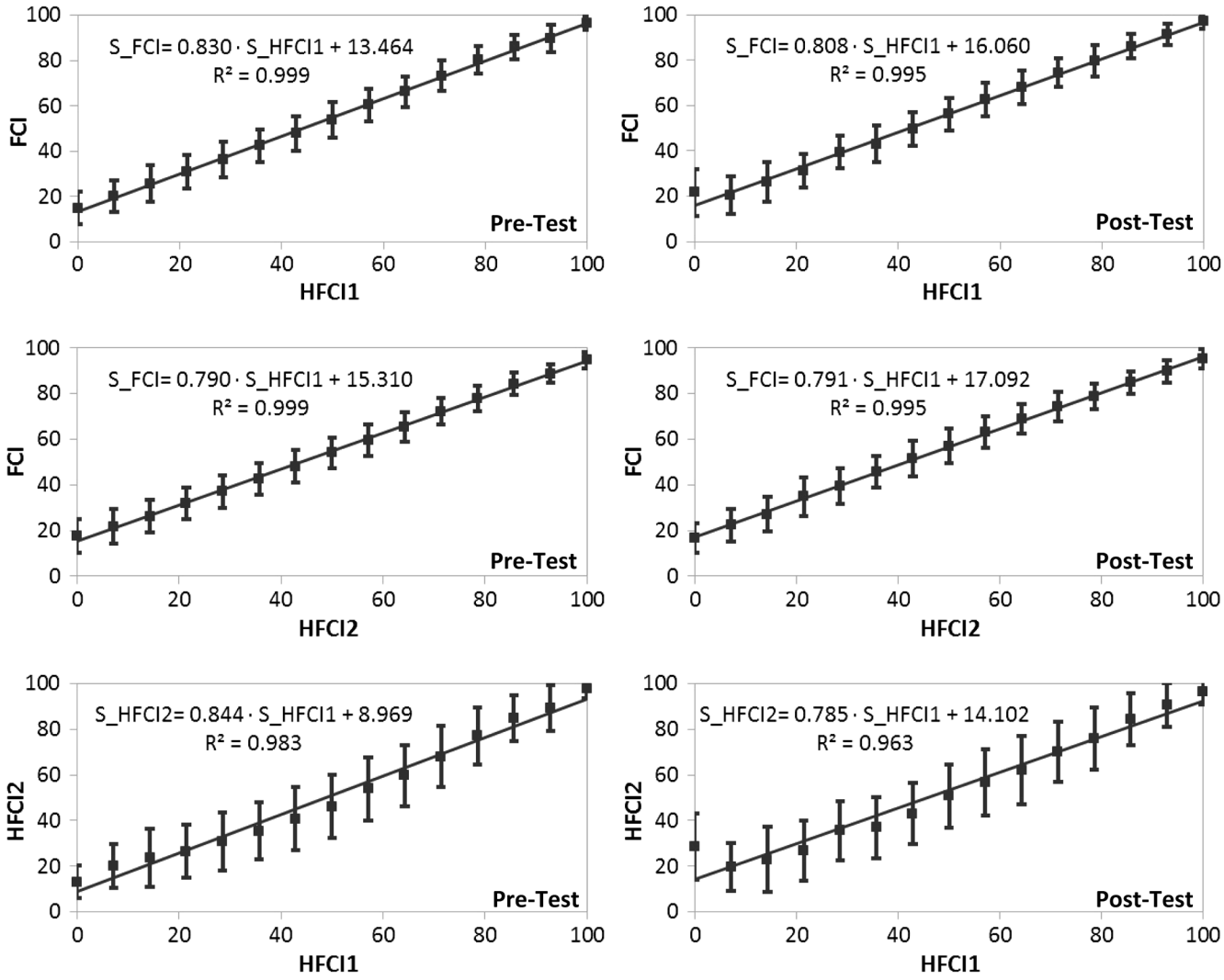

FIG. 2. Cross comparisons between FCI and the half-length tests. In each graph, there are 15 bins along the horizontal axis with an equal difference of $1/14$ of the total score. A point on the graph represents the average score of the test described on the vertical axis from students having a score within a particular bin on the horizontal axis. The error bars are the standard deviations.

error bar of each point gives the standard deviation of the average score of the test indicated on the vertical axis.

The scatter plots show that the binned average scores of the different tests are linearly related at all performance levels. This is an important relation that allows us to use linear functions for score conversions between different tests. Since the making of the short tests are based on the total scores, it is not guaranteed that students at different performance levels will respond equally (linearly) to the new tests. The linear relations shown in Fig. 2 confirm that students at different performance levels respond similarly to both the short version tests and the FCI test. This is also consistent with the similar discrimination parameters of the short tests and the FCI.

The equation for each linear fit is calculated and shown in each of the six figures with R^2 close to 1. (R^2 is the square of the correlation coefficient and gives the fraction of variance explained by the fitted line.) The parameters of the linear fits are listed in Table IV. There are slight variations among the fitting parameters of the different pairs of tests. For consistency purposes, we use the average fitting parameters (also shown in Table IV) to produce two average fitting models in Eqs. (2) and (3), one for converting the scores between a short test and the FCI and the other for converting the scores between the two short tests. Additional conversion models can also be between specific pairs of tests using the results in Table IV.

$$\text{Score\_FCI} = 0.804 \times \text{Score\_HFCI} + 15.482 \quad (2)$$

$$\text{Score\_HFCI2} = 0.814 \times \text{Score\_HFCI1} + 11.536 \quad (3)$$

The uncertainties of score conversions can be evaluated using the differences between the predicted scores and the observed scores with two measures, the mean error (ME) and the root mean square deviation (RMSD), which

TABLE IV. Parameters of the linear fits between any two of the three tests on both pre- and post-tests.

Comparison conditions	Slope	Intercept	R^2
HFCI1 vs FCI for pretest	0.830	13.464	0.999
HFCI2 vs FCI for pretest	0.790	15.310	0.999
HFCI1 vs FCI for post-test	0.808	16.060	0.995
HFCI2 vs FCI for post-test	0.791	17.092	0.995
Average parameters of fitting between FCI and HFCI1/2	0.804	15.482	
HFCI1 vs HFCI2 for pretest	0.844	8.969	0.983
HFCI1 vs HFCI2 for post-test	0.785	14.102	0.963
Average parameters of fitting between HFCI1 and HFCI2	0.814	11.536	

provide quantitative evaluations of the accuracies of score conversions:

$$\mathrm{ME}_{ij} = \frac{\sum_{k=1}^{N} (P_{ij}^{k} - S_{i}^{k})}{N}$$
$$\mathrm{RMSD}_{ij} = \sqrt{\frac{\sum_{k=1}^{N} (P_{ij}^{k} - S_{i}^{k})^2}{N}}. \quad (4)$$

Here, for the kth examinee, S_i^k is the observed score of test i and P_{ij}^k is the predicted score of test i based on the observed score on test j. N is the number of examinees. The errors of score conversions for each pair of the three tests are given in Table V.

As discussed earlier, score conversions can be calculated using individual fitting parameters or the average fitting parameters listed in Table IV. Table V gives the errors of both types of conversions. In general, the mean errors of score conversions are small, with the maximum being 3.49%, which is approximately one-half of a question difference for the short tests. Conversions between the short tests and the FCI have smaller errors than conversions between the two short tests. This is largely due to the smaller numbers of questions in the half-length tests; shorter tests will cause larger steps in total scores, which can lead to larger scales in uncertainties. This is also evident from the larger RMSD measures of the short tests, which double that of the FCI.

The results in Table V also show that using individual fitting parameters rather than the average fitting parameters reduces the error slightly by 0.5%–1.5%. However, since the errors associated with using either method are small, it is preferable to use the average fitting parameters so that score conversions and comparisons in different studies can be more standardized with a consistent conversion model.

From Table V, it appears that by using the average conversion models described in Eqs. (2) and (3), the range of mean errors for predicting equivalent FCI scores from the two half-length test scores on both the pre- and post-test is within ±1.5%, while the RMSD is approximately 7%. The scale of the error is consistent with the theoretically estimated uncertainty based on the IRT response curves shown in Fig. 1. For the FCI test, which has 30 questions, 1.5% error in score is less than the uncertainty produced by half of a question difference. The RMSD is equivalent to the standard deviation of differences between the observed and predicted scores. Based on existing research on control treatment and pre-post comparisons, typical FCI pre-post score differences are larger than 10%, and the standard deviations are on the order of 15%–20%. Therefore, the 1.5% mean error and 7% RMSD provide a good statistical baseline for using the short tests as FCI-equivalent measures.

Meanwhile, using the two half-length tests as parallel tests in short term pre-post testing studies will incur a mean error of approximately 3.5% with an RMSD at the 13.5% level. This range of uncertainty is also small enough for using the two short tests in typical education studies. For a score difference of 10% or more, the error is within 1/3 of the magnitude of the signal, which will not undermine the statistical validity of the study.

TABLE V. Errors between the predicted scores and the observed scores. All scores are in the scale of 0–100.

Prediction conditions	Individual fitting[a]		Average fitting[b]	
	ME	RMSD	ME	RMSD
HFCI1 predict FCI on pretest	0.51	7.39	1.39	7.53
HFCI2 predict FCI on pretest	0.52	6.81	1.33	6.91
HFCI1 predict FCI on post-test	0.09	6.83	−0.70	6.87
HFCI2 predict FCI on post-test	−0.19	6.57	−0.96	6.65
HFCI1 predict HFCI2 on pretest	2.22	13.33	3.49	13.61
HFCI1 predict HFCI2 on post-test	1.01	13.40	0.28	13.28

[a]Errors are evaluated using the individual fitting parameters listed in Table IV.
[b]Errors are evaluated using the average fitting parameters listed in Table IV.

TABLE VI. Comparisons of test performances with data from three different institutions. University 1 is the institution for the original data set. The results for University 1 are copied from Table II for easy comparisons. University 2 ($N = 501$) is a public state university in a different state. The high school data ($N = 657$) are from a suburban high school located in a state different from both universities. The scores and normalized gains in the Table are in the scale of 0–100.

	Pretest			Post-test			Pre-Post score change (ΔS) and normalized gain (g)		
	HFCI1	HFCI2	FCI	HFCI1	HFCI2	FCI	HFCI1	HFCI2	FCI
Testing conditions	Score	Score	Score	Score	Score	Score	ΔS (g)	ΔS (g)	ΔS (g)
University 1 (SD)	43.75 (19.79)	43.66 (21.15)	49.26 (18.17)	62.25 (21.55)	61.93 (23.18)	66.23 (19.22)	18.5 (32.89)	18.27 (32.43)	16.97 (33.45)
University 2 (SD)	30.00 (16.35)	26.86 (16.69)	33.88 (15.10)	41.54 (20.30)	38.71 (21.03)	45.84 (18.64)	11.55 (16.50)	11.84 (16.09)	11.96 (18.02)
High school (SD)	22.29 (13.09)	21.56 (13.48)	25.36 (11.94)	48.95 (24.90)	46.43 (26.52)	51.67 (23.66)	26.66 (34.38)	24.88 (31.80)	26.31 (35.29)

V. TEST EQUATING RELIABILITY

The half-length versions of FCI tests are developed based on a large data set from one university. Since student performances on FCI are dependent on student background and instructional settings, the fact that the data are from a single population can pose potential limitations on the applicability of this research to other student populations with different background and receiving instruction via different pedagogical approaches. To evaluate the reliability of the half-length tests and the equating method, two additional data sets are analyzed and compared together with the original results in Table VI.

The three institutions listed in Table VI are located in three different states. The data set from University 1 is the original data used to develop the short tests. The results for University 1 are copied from Table II for easy comparisons with others. University 2 is a large state university in a different state and the data collected are from an introductory algebra-based college physics course. The third school is a suburban high school with well-established physics courses. The data from University 2 ($N = 501$) and the high school ($N = 657$) are matched pre- and post-test data. With these sample sizes, the standard errors of the average test scores are under 1%.

The three populations and courses compared in Table II are very different, which are confirmed by their pretest scores and pre-post gains. Comparisons of these diverse data sets show that the uncertainty between the half-length tests in measuring different populations in pre- and post-test is approximately 3%. The measurements of score changes are more stable with an overall uncertainty less than 1.5% for all three tests. The uncertainties for normalized gains are on the order of 3%, slightly larger than score changes.

Combining all the analyses discussed above, the reliability of using the short tests with different populations can be initially established at the 3% level. The reliability evaluation is an ongoing process that requires large collections of diverse data sets. The results presented here serve the purpose to provide a starting baseline for applying the half-length tests in future studies. The outcomes of this study suggest that the method of creating short tests based on the FCI instrument can provide FCI-equivalent measurement in the form of total average scores with an overall uncertainty of 3%. Therefore, if the expected score changes or differences are 3 times larger than the uncertainty (>9%), the confidence for trusting the observed signal being a real signal rather than a system error due to using different instruments can be established at a level of 99% or greater. In such cases, the two half-length tests can be used to replace the FCI in assessment.

VI. SUMMARY AND DISCUSSIONS

Conceptual surveying and pre-post testing are valuable methods widely used in education research and teaching practices. When using these methods in practice, researchers and teachers are often concerned about the length of the assessment and memorization from pre- to post-test. If a test is too long, instructors may be reluctant to administer the test because it might take up too much time in an already tightly scheduled classroom. If the same test is used in both pre- and post-testing, at least 5 weeks of time must elapse between the two tests in order to reduce the influence from memorization. In physics education research, the Force Concept Inventory is a popular tool and has established an extensive collection of test results. It would be useful, therefore, to have short, parallel assessments that provide measures equivalent to that of the FCI.

In this paper, we have derived two half-length tests from the popular FCI. The FCI and the two tests proposed here, extracted as subsets of the FCI, cover the same concepts and appear to have similar assessment characteristics that produce equivalent test scores with an overall uncertainty less than 3%. With these new parallel tests, researchers and teachers would be able to administer the assessment in a shorter period of time, reduce the test-retest effects in pre-post testing, and still expect to measure scores directly comparable to that of the full FCI test. Although the outcomes of this study are encouraging, the new tests

and equating methods will need to be thoroughly validated through additional research involving a large number of diverse data sets from different populations and institutions.

The nature of this study represents a theoretical data-mining approach on modifying an existing instrument. There are limitations and constraints. The assessment method can produce FCI-equivalent mean scores using two half-length tests. The results only apply to the mean scores and should not be used to make inference about student understanding of specific concepts as a shortened test will certainly leave out a great deal of assessment capacity. Therefore, the methods and results introduced in this paper are only applicable to average total test scores and should not be used to evaluate scores of individual students or subsets of the questions on specific concepts included in the tests.

In addition, the data-mining technique uses the existing data collected with the full 30-question FCI. An experimental study that uses all three tests with randomized samples of identical populations has been carried out to further establish the reliability of the short version tests. The preliminary results are consistent with the theoretical predictions, which will be reported in a follow-up paper.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Physics Education Research Group at The Ohio State University for their long-term support and valuable suggestions on this project. We would also like to acknowledge the help of the two anonymous reviewers. This research is supported in part by NIH Grant No. RC1RR028402 and NSF Grants No. DUE-0633473 and DUE-1044724.

[1] D. Hestens, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] 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).
[3] X. Liu and D. MacIsaac, An Investigation of Factors Affecting the Degree of Naive Impetus Theory Application, J. Sci. Educ. Technol. **14**, 101 (2005).
[4] I. A. Halloun and D. Hestenes, Common sense concepts about motion, Am. J. Phys. **53**, 1056 (1985).
[5] S. F. Itza-Ortiz, S. Rebello, and D. Zollman, Students' Models of Newton's Second Law in Mechanics and Electromagnetism, Eur. J. Phys. **25**, 81 (2004).
[6] K. Hill and A. Wigfield, Test Anxiety: A Major Educational Problem and What Can Be Done About It, Elementary School J. **85**, 105 (1984).
[7] S. Krause, J. Birk, R. Bauer, B. Jenkins, and M. J. Pavelich, Development, Testing, and Application of a Chemistry Concept Inventory, in *34th ASEE/IEEE Frontiers in Education Conference*, Savannah, GA, 2004, 0-7803-8552-7/04.
[8] C. Henderson, Common Concerns About the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).
[9] M. E. Otter, G. J. Mellenbergh, and K. de Glopper, The Relation between Information-Processing Variables and Test-Retest Stability for Questionnaire Items, J. Educ. Measure. **32**, 199 (1995).
[10] J. S. Smith, D. Y. Dai, and B. P. Szelest, Helping First-Year Students Make the Transition to College through Advisor-Researcher Collaboration, NACADA J. **26**, 67 (2006).
[11] J. Kulik, C. C. Kulik, and R. L. Bangert, Effects of Practice on Aptitude and Achievement Test Scores, Am. Educ. Res. J. **21**, 435 (1984).
[12] P. Heller and D. Huffman, Interpreting the Force Concept Inventory. A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[13] L. Bao, Ph.D. Dissertation, University of Maryland, 1999.
[14] L. Bao and E. F. Redish, Model Analysis: Assessing the Dynamics of Student Learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).
[15] J. Stewart, H. Griffin, and G. Stewart, Context sensitivity in the force concept inventory, Phys. Rev. ST Phys. Educ. Res. **3**, 010102 (2007).
[16] M. J. Kolen and R. L. Brennan, *Test Equating* (Springer-Verlag, New York, 1995).
[17] F. Baker, Ability metric transformations involved in vertical equating under item response theory, Appl. Psychol. Meas. **8**, 261 (1984).
[18] J. Wang and L. Bao, Analyzing Force Concept Inventory with Item Response Theory, Am. J. Phys. **78**, 1064 (2010).
[19] L. Chen, J. Han, J. Wang, Y. Tu, and L. Bao, Comparisons of Item Response Theory Algorithms on Force Concept Inventory, *Res. Educ. Assess. Learn.* **2**, 26 (2011).
[20] M. B. Wilk and R. Gnanadesikan, Probability plotting methods for the analysis of data, Biometrika (Biometrika Trust) **55**, 1 (1968).
[21] For people interested in IRT algorithms, it is noticed that the test parameters in Table III are not the traditionally defined item parameters. In IRT, the response parameters of a test have not been well established. Our method in taking the averages of item parameters of individual test items is a simple first-order approximation for comparing the different tests on their overall test item characteristics. As a result, the test parameters in Table III do not have simple relations to the total test scores, although individual test items often do.