Experimental validation of the half-length Force Concept Inventory

Jing Han,[1] Kathleen Koenig,[2] Lili Cui,[3] Joseph Fritchman,[1] Dan Li,[4] Wanyi Sun,[1] Zhao Fu,[1] and Lei Bao[1,*]

[1]*The Ohio State University, Columbus, Ohio 43210, USA*
[2]*University of Cincinnati, Cincinnati, Ohio 45220, USA*
[3]*North Carolina State University, Raleigh, North Carolina 27695, USA*
[4]*Beijing Jiaotong University, 100044, Beijing, China*

(Received 7 July 2015; revised manuscript received 2 June 2016; published 4 August 2016)

In a recent study, the 30-question Force Concept Inventory (FCI) was theoretically split into two 14-question "half-length" tests (HFCIs) covering the same set of concepts and producing mean scores that can be equated to those of the original FCI. The HFCIs require less administration time and reduce test-retest issues when different versions are used in pre-post testing. This study experimentally evaluates the practical validity and measurement uncertainty of the HFCIs with three different college student populations. Measured mean scores on each HFCI were within ~3% of each other at every university. Measured mean score differences between the HFCI and FCI were also within ~3%. These differences are less than the value of a single question on the 30-question FCI and are not statistically significant. The overall results suggest that, in conditions similar to this study, the HFCIs can be used as alternatives to the full-length FCI when total scores or score gains are the measurement goals.

DOI: 10.1103/PhysRevPhysEducRes.12.020122

I. INTRODUCTION

Beginning with the early successes of concept inventories (CIs), such as the Force Concept Inventory (FCI) [1], many CIs have been developed to probe student understanding of fundamental concepts in science, technology, engineering, and mathematics (STEM) classrooms [2]. The FCI is the most commonly used assessment in physics education, resulting in major advancements such as demonstrating the effectiveness of interactive engagement methods with the Hake study [3]. Part of the effectiveness of the FCI is due to its large available data sets, which provide excellent opportunities for researchers to develop new analysis techniques and compare these to understood results [4,5].

While the FCI has worked well, time constraints and test-retest memorization issues often limit the use of the FCI and other CIs in teaching and research [6–10]. The need for shorter CIs motivated the Han *et al.* study [11], which used existing data sets and data mining and modeling methods to develop half-length versions of the FCI (HFCI). These two half-length FCIs (HFCI1 and HFCI2) were created with the goal of being equivalent parallel tests that have similar assessment capabilities to each other and to the original FCI. Questions for each version were chosen based on contexts, average scores, and item response theory estimates of assessment features. If proven effective, similar methods may be used when creating or shortening other CIs.

The HFCIs were developed as a way to more efficiently administer CIs. In order for the HFCIs to be considered equivalent, each version would need to produce equivalent scores and assessment characteristics. Modeled and evaluated with the existing data, the total scores of the two HFCIs developed in the Han *et al.* study [11] were found to be virtually identical (within the 0.5% standard error on the pretests and 0.6% standard error on the posttests), and about 5% lower than the FCI scores [11]. The lower scores on the HFCI were expected due to removal of several higher-scoring FCI questions (on the HFCI). The differences between HFCIs and FCI were found to be consistent with data sets from different populations. This allows the equating of HFCIs scores and FCI scores using a linear conversion model that translates scores on one test to scores on the other. Altogether, the total scores, score changes, and normalized gains of the two short tests provide equivalent measures of student performances and learning gains [11].

It is important to note that the method in Han *et al.* [11] is aimed to produce equivalent total scores and, therefore, the results of the shortened tests apply only to the total scores and shall not be used to draw conclusions regarding student understanding of specific concepts. Sensitivity to context often limits the capacity of the shortened tests in measuring subcategories of conceptual understanding.

The real HFCIs now need to be tested with actual students to determine how close the results of the short tests and the original FCI match in practical applications

*bao.15@osu.edu

and what size of errors may be expected. This paper applies the HFCIs to samples from three different populations to analyze the usability of the tests in practice. Specifically, the following two research questions will be investigated:

(1) Do the two HFCIs produce equivalent score measures in different populations and what error sizes should be expected?
(2) How do the results of the HFCIs compare with the FCI results?

II. METHOD AND DESIGN

Based on the theoretical work [11], the actual HFCI tests are created. The question order of the HFCIs and the mapping to the original FCI test are included in Table I. Question 11 (FCI question 26) of the HFCI1 is the second question in a sequence on the FCI-95. Therefore, it has been slightly edited to include the narrative of a referring question (Q25) so that it can be a standalone question.

In education assessment, measured scores of different tests are influenced by both the features of the tested samples and the tests themselves. In this study, three questions are kept common to all versions of the tests (FCI questions 2, 26 and 28; see Table I). These are also placed at identical positions in the HFCIs to reduce potential influences due to question positions in the tests. This design, which allows the comparisons of students' scores on common and different questions on the different tests, provides additional information to analyze the extent to which score differences of different tests may be influenced from students' variations and/or test constructs.

TABLE I. FCI to HFCI mapping. The question order for each HFCI is shown below, with mapping to FCI-95 questions. Questions in bold font appear on both versions of the HFCI and in the same question position. Question 11 in HFCI1, with an asterisk (*), was slightly edited to include the narrative of the referring question (Q25) to make it suitable as a standalone question. To obtain the actual test, see the Table footnote[a].

HFCI question order	HFCI1-FCI mapping	HFCI2-FCI mapping
1	**2**	**2**
2	15	4
3	5	30
4	6	7
5	8	18
6	9	21
7	10	22
8	11	23
9	13	24
10	12	25
11	**26***	**26**
12	17	14
13	19	20
14	**28**	**28**

[a]To protect the validity of the tests, the HFCIs are not made publicly available. However, all teachers who wish to use the tests may contact the authors to access the tests.

The data collection uses all three versions of the tests, including the original FCI-95 and the two HFCIs. The subjects include students enrolled in introductory physics courses from three universities in three different states. Each university (A, B, and C) administered the tests in recitations of the courses that were taught using traditional lecture methods (see Table II). The version of the test a student took was randomized by interweaving stacks of identical numbers of the different tests before handing them out to students.

University A's sample consisted of students in calculus-based introductory mechanics courses during the Fall 2012 ($N = 197$) and Spring 2013 ($N = 188$) semesters. These students were enrolled in science and engineering programs at the university. The tests were administered during the middle of the consecutive semesters. Only the two HFCIs were tested at this location.

University B's sample ($N = 102$) consisted of students in an algebra-based introductory mechanics course during the Fall 2012 semester. The tests were administered as a pretest. Only the two HFCIs were tested at this location.

University C's sample ($N = 913$) consisted of students in a calculus-based introductory mechanics course during the Fall 2012 semester. These students were enrolled in science and engineering programs. Because of the large sample size at this university, the FCI-95 was also administered. To help control for test length, additional filler questions were added after the HFCIs so that the time requirements were similar to the full FCI. These tests were all administered as a posttest. It is also noted that the HFCI tests were developed using previous FCI data from this university [11].

If the HFCIs work as they are intended, one university's mean score on each HFCI should be equivalent to that university's mean score on the other HFCI and the mean scores may differ institution to institution. Analyzing this pattern in different populations will provide evidence to demonstrate the extent of the reliability of the HFCIs. In

TABLE II. Sample sizes for each university in this study. Total sample sizes for each university are given, as well as the breakdown for how many students completed each version of the FCI. Students from universities A and C were enrolled in science and engineering programs, while students from university B were enrolled in a teaching program. Only university C tested the full FCI. (Pre: administered as pretest, Mid: administered during the middle of the semester, and Post: administered as posttests).

University (Time frame)	Time	Total N	N for HFCI1	N for HFCI2	N for full FCI
A (mid)	Fall 2012	197	104	93	
A (mid)	Spring 2013	185	92	96	
B (pre)	Fall 2012	102	51	51	
C (post)	Fall 2012	913	289	295	329

addition, the differences and conversions between HFCIs and the FCI-95 can also be analyzed to inspect if they are consistent with the theoretical predictions [11].

III. RESULTS AND DISCUSSION

The data collected from the two HFCIs and the FCI-95 were compared to answer the two research questions. The main variable of the analysis was the mean scores of the different tests from samples of different populations. The statistical analysis used analysis of variance (ANOVA), student T-test, and effect sizes to compare the mean scores and differences of mean scores.

HFCI scores at universities A, B, and C are recorded in Table III, and show small differences consistent within statistical uncertainty (1%–3%). These differences correspond to scores with less than one question difference on the original FCI. It is worth noting that by increasing the sample sizes, the standard error of a sample's mean score is reduced. Therefore, with large samples, even small differences can be statistically significant. In many cases, the effect sizes will be a good way to evaluate if a difference is meaningful. In typical education assessment, effect sizes smaller than 0.2 are often considered small and insignificant for educational purposes.

In this study, the sample sizes of the different test conditions range from 50 to 300, which are typical class sizes of college courses. In all these cases, the differences are not statistically significant ($p \gg 0.05$) and the effect sizes range from 0.0 to 0.14. These results provide the experimental basis to answer the first research question, suggesting that the HFCIs are suitable for use as equivalent tests in college courses of class sizes ($N > 50$) and populations similar to the ones in this study.

In addition, students' scores on common questions were analyzed to provide further information on the extent to which the score difference between the two tests may be caused by the variations of the students and the possible influences from test constructs, such as question orders and the length of the test. In theory, if two tests have identical assessment features, one should expect nearly identical score differences on common and noncommon questions. The results in Table III show that the total scores and the scores of the common questions are in the same range; the differences are not statistically significant.

University C was the only location to test students using both HFCIs and the full FCI-95, with scores shown in Table IV. All the measured scores on the HFCIs and FCI-95 are slightly lower than the predictions of Han *et al.* [11], where the measured average HFCI score is 3.47% lower than predicted and the measured FCI-95 score is 4.87% lower than predicted. These differences are often expected due to student variations from year to year. Comparing the two HFCIs, the measured score difference is 0.73%, which is slightly larger than the predicted value of 0.32%, but still represents a good agreement between the two tests ($p = 0.706$ and effect size $= 0.031$; see Table III).

Comparing the FCI-95 and the HFCIs, the measured score differences between the FCI-95 and the HFCIs follow the predicted trend, showing that the FCI-95 produces slightly higher scores than the HFCIs due to the removal of five easier questions [11]. If the five questions that were not used on either HFCI are removed from scoring on the FCI,

TABLE III. Mean HFCI scores at universities A and B. HFCI scores at universities A and B are given as well as scores on the three questions that the HFCI versions had in common. Score differences (ΔS = HFCI2−HFCI1) and their p values and effect sizes are given to the right. The average score is a nonweighted average of the sample means. Standard error (SE) and sample size (N) are listed for each score.

	Tests		Common questions		ΔHFCI1 and HFCI2	
University	HFCI1	HFCI2	HFCI1	HFCI2	All	Common
A (Fall 2012)	54.81	52.53	38.78	40.50	$\Delta S = -2.27$	$\Delta S = 1.72$
(SE)	(2.14)	(2.64)	(3.06)	(3.49)	$p = 0.504$	$p = 0.711$
(N)	(104)	(93)	(104)	(93)	eff $= 0.096$	eff $= 0.053$
A (Spring 2013)	53.34	56.62	45.65	50.35	$\Delta S = 3.28$	$\Delta S = 4.70$
(SE)	(2.46)	(2.64)	(3.12)	(3.36)	$p = 0.340$	$p = 0.270$
(N)	(92)	(96)	(92)	(96)	eff $= 0.141$	eff $= 0.163$
B (Fall 2012)	27.87	28.01	32.67	32.67	$\Delta S = 0.14$	$\Delta S = 0.00$
(SE)	(1.64)	(1.40)	(4.43)	(4.01)	$p = 0.948$	$p = 1.0$
(N)	(51)	(51)	(51)	(51)	eff $= 0.013$	eff $= 0.00$
C (Fall 2012)	58.26	58.98	50.52	51.07	$\Delta S = 0.73$	$\Delta S = 0.55$
(SE)	(1.33)	(1.40)	(1.82)	(2.00)	$p = 0.706$	$p = 0.838$
(N)	(289)	(295)	(289)	(295)	eff $= 0.031$	eff $= 0.017$
Average Score	48.57	49.04	41.91	43.65	$\Delta S = 0.47$	$\Delta S = 1.74$

TABLE IV. Mean HFCI and FCI scores at university C. Scores for each test version are given: HFCI1, HFCI2, FCI-25Q (FCI-95 with five questions not used on either HFCI omitted), FCI-30Q (full FCI-95). FCI-30 scores are slightly higher than the HFCI scores, but if the five questions not used in either HFCI are omitted, the FCI-25 score falls in between the two HFCI scores. Scores on the three questions (FCI questions 2, 26, and 28) common to all versions of the test are higher in the FCI than in either HFCI. Standard error (SE) and sample size (N) are listed for each score.

	Tests				Common questions		
	HFCI1	HFCI2	FCI-25Q	FCI-30Q	HFCI1	HFCI2	FCI-95
University	($N = 289$)	($N = 295$)	($N = 329$)	($N = 329$)	($N = 289$)	($N = 295$)	($N = 329$)
C (Fall 2012)	58.26	58.98	58.46	61.36	50.52	51.07	53.70
(SE)	(1.33)	(1.40)	(1.27)	(1.20)	(1.82)	(2.00)	(1.87)

the FCI score shifts to 58.46%, which falls in between the two HFCI scores (average of 58.62%) and the effect sizes are minimal when compared to the HFCIs. The result may suggest that the 25-question version of the FCI scales better with the HFCIs than does the full FCI-95. However, explorations on using the 25-question version FCI in practice is a valuable topic but is beyond the scope of this paper.

When comparing the common questions between FCI-95 and HFCIs, FCI-95 scored 2.9% higher than the HFCIs did on average. Between the HFCIs, the difference on common questions is just 0.55%, equivalent to an effect size of 0.017. The difference between mean scores of FCI-95 and HFCIs is not statistically significant as ANOVA results show [$F(2910) = 1.59$, $p = 0.20$], and corresponds to an effect size of 0.122. In addition, differences in the scores on common questions are also not significant [$F(2910) = 0.83$, $p = 0.44$]. While this study does not support a conclusive claim, it may be worthwhile to further examine these common questions. If a difference was found to exist with a larger sample size, it would have implications regarding contextual differences on students' performances between the long and short tests.

However, the differences are on the scale of 3%, which gives an average effect size less than 0.15 in typical experimental cases. A 3% difference on three questions on the FCI would average to a 0.3% difference on the entire test, while it averages to a 0.6% difference on the HFCIs. These differences are less than the uncertainty on each test. As a result, it is reasonable to conclude that the contextual influences due to shortening the FCI-95 can be ignored in typical testing conditions.

In general, identical questions placed in different tests will often result in variations in their measured scores. To explore the extent of such variations when the identical sets of questions appear in the HFCIs and the full FCI, the scores for the full FCI were broken down into scores on questions appearing on each version of the HFCIs (labeled as HFCIs*) and compared with the scores of the HFCIs as standalone tests. The results are summarized in Table V, and show that the differences between any of the HFCIs, HFCIs*, and the FCI-25Q are below 2%. A repeated measures ANOVA demonstrates that the scores on each portion of the breakdown are not statistically different [$F(2328) = 1.94$, $p = 0.15$]. The analysis suggests that the differences in test length and contextual configurations of the HFCIs and the full FCI do not alter the measured scores statistically. This comparison further supports the claim that the HFCIs and the FCI can be considered equivalent in score-based measurements.

Han *et al.* also proposed conversion models to translate between HFCI scores and from HFCI scores into FCI-95 scores [11], which are useful for making comparisons between different test scores and with historical FCI data.

First, the models for converting scores between the two HFCIs are tested. In Table VI, the measured HFCI scores and the predicted HFCI2 scores are listed. The HFCI2 predicted scores are converted from the measured HFCI1 scores using the linear conversion models in Han *et al.* [11]. Based on the time frame of the measurement, the pretest results are calculated with the pretest conversion model while the posttest results are calculated with the posttest conversion model. If a test was administered in the middle of the instruction (e.g., for university A), the average

TABLE V. Mean FCI scores at university C broken down into scores on questions appearing on each version of the HFCI, marked with an asterisk (*). Unmarked are the actual test data copied from Table IV. Standard error (SE) and sample size (N) are listed for each score.

University	HFCI1* ($N = 329$)	HFCI1 ($N = 289$)	HFCI2* ($N = 329$)	HFCI2 ($N = 295$)	FCI-25Q ($N = 329$)	FCI-30Q ($N = 329$)
C (Fall 2012)	58.77	58.26	57.12	58.98	58.46	61.36
(SE)	(1.32)	(1.33)	(1.37)	(1.40)	(1.27)	(1.20)

TABLE VI. Measured HFCI scores and predicted HFCI2 scores. The HFCI2 predicted scores are obtained by applying the linear conversion model to the measured HFCI1 scores. The differences between predicted and measured HFCI2 scores are also calculated and listed to the right. (Pre: administered as pretest, Mid: administered during the middle of the semester, and Post: administered as posttests).

University (Time frame)	HFCI1 (Measured)	HFCI2 (Measured)	HFCI2 (Predicted)	ΔHFCI2 (Predicted-Measured)
A (mid)	54.81	52.53	56.15	3.62
A (mid)	53.34	56.62	54.95	−1.67
B (pre)	27.87	28.01	32.49	4.48
C (post)	58.26	58.98	59.84	0.86

conversion model was used. The differences between the measured and predicted scores are also computed. In general, the differences between measured and predicted scores are consistent with the computed uncertainties of the conversion models, which have a mean error typically around 3% [11].

The results in university C appear to provide the best match. It is worth noting that the conversion models were developed based on historical data from university C. Therefore, it can be inferred that the parameters of the conversion model are dependent on the population backgrounds; however, a conclusive claim on the extent of this dependence will need further study with a larger set of different populations.

Since university C provides scores of all three versions of tests, the data can be used to experimentally evaluate the conversion models that predict scores of the full FCI from either of the HFCIs. The results are listed and compared in Table VII. The predicted FCI scores from both HFCIs are slightly higher than the measured score. The differences are around 2%, well below the estimated uncertainty of the conversion models, which is around 7% [11]. The difference between the two predicted FCI scores is 0.62%, suggesting that the conversion models from both HFCI tests are consistent.

In physics education research, the normalized gain (g) is a popular measure widely used for evaluating pre-post changes [3]. It would be useful to estimate the changes to the uncertainty in g when switching from the FCI to the HFCIs and to determine under what conditions it would be feasible to use HFCIs with the normalized gain. There are several versions of calculations for g [12]. Here the discussion focuses on the formulation that uses the average pre- and posttest scores. Denoting $\bar{x}$ as the average pretest score and $\bar{y}$ as the average posttest score and assuming $\bar{y} \geq \bar{x}$, the normalized gain has the following form:

$$g = \frac{\bar{y} - \bar{x}}{1 - \bar{x}}. \tag{1}$$

The maximum uncertainty in the normalized gain as a result of the measurement uncertainties in the average pre- and posttest scores can then be obtained from

$$\delta g \leq \left|\frac{\partial g}{\partial \bar{x}}\right| \delta \bar{x} + \left|\frac{\partial g}{\partial \bar{y}}\right| \delta \bar{y}. \tag{2}$$

Here, $\delta\bar{x}$ and $\delta\bar{y}$ are the measurement uncertainties of the average pre- and posttest scores, and δg is the uncertainty of the normalized gain due to the measurement uncertainties of the average pre- and posttest scores. The measurement uncertainty of an average score is usually evaluated with the standard error, which is equal to the standard deviation divided by the square root of the sample size. Based on data, the typical standard deviations for HFCI pre- and posttest scores are similar and have a value around 0.22. For simplicity, a single standard deviation σ is used to present the uncertainty in both pre- and posttest scores. Assuming that the sample sizes for the pre- and posttest are also identical, the maximum uncertainty of g can be simplified as

$$\delta g \leq \frac{2 - (\bar{x} + \bar{y})}{(1 - \bar{x})^2} \frac{\sigma}{\sqrt{N}}. \tag{3}$$

The maximum uncertainty varies with the uncertainty of the score and the inverse square root of the sample size. When treating each question in a test as an independent measure, the uncertainty in the average score of the test should vary with the inverse square root of the number of questions (i.e., half the questions would lead to an uncertainty $\sqrt{2}$ larger). The typical standard deviation for HFCI scores is around 0.22, while the typical standard deviation for FCI scores is around 0.15.

In Table VIII, simulated estimates using Eq. (3) for the maximum uncertainty of g at various class sizes are given for the FCI and HFCI. When the class size is relatively large ($N > 100$), the difference of uncertainties calculated with FCI and HFCI is very small, about 0.015. With smaller class sizes ($N < 30$), the uncertainty of g calculated from HFCI scores can get close to 0.1, about 0.03 larger than the uncertainty when the full FCI scores are used. Such cases may create issues comparing between traditional and interactive engagement classrooms (where the typical difference of g is around 0.2). Therefore, it appears that while the HFCI should work well with class sizes similar to

TABLE VII. Measured HFCI and FCI scores and predicted FCI scores. The predicted FCI scores are obtained by applying the linear conversions to the two measured HFCI scores at university C. The differences between the predicted FCI scores (from HFCI1 and HFCI2) and the measured FCI score are calculated and listed as Δ1 and Δ2. The difference between the two predicted FCI scores are listed as Δ3.

HFCI1 Measured	HFCI2 Measured	FCI Measured	FCI from HFCI1	FCI from HFCI2	Δ1	Δ2	Δ3
58.26	58.98	61.36	63.13	63.75	1.77	2.39	0.62

TABLE VIII. Benchmark estimates for the uncertainty in the normalized gain for FCI and HFCI. The estimates are calculated by assuming that $\bar{x} = 0.3$, $g = 0.4(\bar{y} = 0.58)$, as often expected from interactive engagement classrooms, $\sigma_{\mathrm{HFCI}} = .22$, and $\sigma_{\mathrm{FCI}} = .15$.

N	δg (FCI)	δg (HFCI)
30	0.063	0.092
50	0.048	0.071
100	0.034	0.050
300	0.020	0.029

those used in this study ($N > 50$), the uncertainty in normalized gains may be too large in smaller class sizes ($N < 30$).

Finally, it is worth noting a limitation of this study. There is only one data set from university C that has used all three versions of the tests. Therefore, the conversion model from HFCIs to FCI has only been tested once (Table VII). From the testing of the conversions between HFCIs (Table VI), the best result was observed at university C, where the conversion models were created using historical data. The relative success of the conversion models at university C compared with the larger discrepancies at the other universities suggests that these models may be dependent on the tested populations and their instruction, and may require fine-tuning with the tested populations. Therefore, it is recommended that if conversion is desired, a set of modified parameters in the conversion equations should be calibrated with the data from the target population. This can be done by administering the full FCI to a portion of students and calculating linear fits between these students' scores on the whole test and the subscores for questions on each of the two HFCIs. See Han *et al.* [11] for more detailed information on the creation of these models.

IV. SUMMARY AND CONCLUSIONS

The half-length FCIs provide a solution to a number of issues with inventory testing. Containing only 14 questions significantly reduces the time for testing, which makes it easier for instructors to implement the assessment. As the two HFCIs are score equivalent, they can be used for short-term pre- and posttest administration to minimize possible item memorization effects.

However, use of the HFCIs will introduce limitations. The content of the test contains fewer context cases, which constrains the assessment of student understanding of specific concepts. While this behavior may not be appropriate for all testing situations, for the majority of studies mainly interested in measuring total scores, the limitations will be inconsequential.

This study demonstrates experimentally that the two half-length FCI tests proposed by Han *et al.* [11] produce equivalent results in different populations. Differences between scores on the two tests were less than the value of one question on the original FCI and are statistically insignificant with minimal effect sizes. These results support Han *et al.* [11] goal of creating equivalent short FCIs. However, caution is advised when using the HFCIs with class sizes smaller than 30 students, as the uncertainty in normalized gains may become too large to compare traditional and active learning classes. If a study is primarily concerned with total scores, no such warning is necessary.

In addition, linear score conversions applied to HFCI1 to predict HFCI2 scores produce errors less than 5%, while the conversions from HFCIs to FCI produce errors around 2%. In most education studies, the expected score gains during an introductory physics course is typically 15% and above, which makes it suitable to use the HFCIs to replace the FCI as equivalent tests. While acceptable for this use, conversion formulas tailored to each institution may provide greater accuracy.

Overall, this study supports treating the HFCIs as interchangeable equivalents to the FCI when average scores and normalized gains are the primary interest. Use of these instruments requires roughly half the time of the original FCI, providing a more efficient method of testing teaching methods in introductory physics courses.

ACKNOWLEDGMENTS

This research is supported in part by NSF Awards No. DUE-1044724, No. NSF-DUE-1431908, and No. NSF-DRL-1417983. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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

[2] D. L. Evans, G. L. Gray, S. Krause, J. Martin, C. Midkiff, B. M. Notaros, D. R. Pavelich, D. Rancour, T. R. Rhoads, P. Steif, R. A. Streveler, and K. Wage, Progress on concept inventory assessment tools, in *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO, 2003* (IEEE, New York, 2003), Vol. 1, pp. 1–8.

[3] 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).

[4] L. Bao and E. F. Redish, Concentration analysis: A quantitative assessment of student states, Am. J. Phys. **69**, S45 (2001).

[5] L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, Phys. Rev. ST Phys. Educ. Res. **2**, 010103 (2006).

[6] K. Hill and A. Wigfield, Test anxiety: A major educational problem and what can be done about it, Elem. Sch. J. **85**, 105 (1984).

[7] C. Henderson, Common concerns about the Force Concept Inventory, Phys. Teach. **40**, 542 (2002).

[8] 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).

[9] 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).

[10] 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).

[11] 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).

[12] L. Bao, Theoretical comparisons of average normalized gain calculations, Am. J. Phys. **74**, 917 (2006).