Taking an Item-Level Approach to Measuring Change With the Force and Motion Conceptual Evaluation: An Application of Item Response Theory

Robert M. Talbot III
University of Colorado Denver

In order to evaluate the effectiveness of curricular or instructional innovations, researchers often attempt to measure change in students' conceptual understanding of the target subject matter. The measurement of change is therefore a critical endeavor. Often, this is accomplished through pre–post testing using an assessment such as a concept inventory, and aggregate test scores are compared from pre to post-test in order to characterize gains. These comparisons of raw or normalized scores are most often made under the assumptions of Classical Test Theory (CTT). This study argues that measuring change at the item level (rather than the person level) on the Force and Motion Conceptual Evaluation (FMCE) can provide a more detailed insight into the observed change in students' Newtonian thinking. Further, such an approach is more warranted under the assumptions of Item Response Theory (IRT). In comparing item-level measures of change under CTT and IRT measurement models, it was found that the inferences drawn from each analysis are similar, but those derived from IRT modeling stand on a stronger foundation statistically. Second, the IRT approach leads to analyzing common item groupings which provide further information about change at the item and topic level.

The measurement of change is necessary for evaluating the effectiveness of instructional innovations in educational contexts. Without measures of change in students' conceptual understanding, we lack strong foundations for making inferences about existing and reform-oriented instructional strategies and curricula. Although measures alone are not sufficient for making these inferences, they are a necessary part of such research. However, measurement itself is difficult. It is part art and part science. There is no single test score that can unequivocally tell us everything we need to know about students' conceptual understanding. Moreover, attempting to measure change in conceptual understanding presents its own problems. Educational researchers have been tackling these problems for years (cf. Cronbach & Furby, 1970; Willett, 1988–89), and lively discussions are still taking place regarding the measurement of change. Challenging as it is, measurement of change must be undertaken at all levels of instruction and across all subjects. In our current educational culture of accountability it is important for us to attempt to measure change in students' conceptual understanding.

In science instruction, concept inventories are often administered to students pre- and post-instruction in order to characterize change in conceptual understanding. The Force and Motion Conceptual Evaluation (FMCE; Thornton & Sokoloff, 1998) is one such concept inventory. It is often used in introductory physics courses to evaluate students' ability to think in a Newtonian fashion. The pre/post administration of this and other concept tests, such as the Force Concept Inventory (FCI) (Hestenes, Wells, & Swackhammer, 1992), is used in physics courses in order to provide evidence for making inferences about changes in students' ability to think in Newtonian terms. These measured changes are then taken to be indicators of course efficacy. The physics education research (PER) community has been using these types of assessments for a number of years (e.g., Bonham, Deardorff, & Beichner, 2003; Finkelstein & Pollock, 2005; Meltzer, 2002; Pollock, 2004; Smith & Wittmann, 2007; Van Domelen & Van Heuvelen, 2002). Assessment work by PER researchers has provided us with a wealth of data to analyze (e.g., Hake, 1998) and has contributed much to the literature on learning in introductory physics courses (e.g., Hake, 2002), specifically with regards to comparing "traditional" approaches to teaching to more interactive or innovative approaches. For example, Bonham et al. (2003) used the FMCE (in addition to other measures) to compare student learning between two groups: one which engaged in paper-based homework assignments and another which engaged with web-based homework.

Discussions about the use of concept inventories in PER take place frequently in various communities and on listservs such as the Physics Learning Research List (PhysLrnR[1]). Though concept inventory use is widespread, it is not without some debate. For example, on many concept inventories, students "hit the ceiling" on the post-test (i.e., obtain a perfect score). This becomes a potential issue when calculating a gain score. Another issue that has

been discussed recently relates to the context in which students' understanding is measured. Are measures of Newtonian Thinking that derive from scores on a concept inventory different from those that might derive from a performance based setting (e.g., lab)? In other words, researchers are thinking critically about the strengths and limitations of these concept inventories. Despite these potential issues, the use of concept inventories provides a sort of common ground upon which we can communicate and compare our findings, and as such their use remains popular in discipline-based education research, especially in PER. This article reports on research which has implications for the "ceiling" issue while accepting the contextual limitations of this and other concept inventories.

Measures of change in PER are most commonly based on a comparison of raw scores from pre–post testing. A student's composite score on an exam serves as a proxy for their ability with respect to the construct of interest, in this case Newtonian thinking. Within that framework, differences in post- and pre-test raw scores are often normalized and used as indicators of the amount of change in student conceptual understanding that has occurred during instruction (e.g., Hake, 1998). Although these normalized raw score difference measures are quite useful as indicators of change in understanding, they have some problems. The measurement models applied to these types of analyses are a part of Classical Test Theory (CTT), which is based on observed raw scores and considers those scores to be composed of true score and error score components. The most notable issues with these CTT measures of change include the raw score bias (and resulting problems in scale), potential low reliability of change scores if correlation between pre- and post-test scores is high, and spurious relationships between gain scores and initial scores due to measurement error (Bereiter, 1963).

This particular study addresses the following research questions: At the item-level, how does an Item Response Theory (IRT) approach to the measurement of change on the FMCE compare to CTT measures of change? Further, do these different approaches to measuring change on the FMCE lead us to make different inferences about student learning of Newtonian Physics?

Item Response Theory is a theoretical approach to designing, analyzing, and scoring tests. It is often associated with current research and work in construct-based measurement (Wilson, 2005) which pays particular attention to the construct as the theoretical object of interest. IRT statistical models are probabilistic models and as such generate estimates of a respondent's ability or the difficulty of items on a test. Because of the rigorous development process and strong statistical foundations, these methods are often used in the development and analysis of high-stakes tests such as the Graduate Record Exam (GRE). However, also because of the characteristics, IRT-based development and analyses are intensive, difficult to learn and carry out, and results are not always straightforward to interpret.

The next section of this article will provide some background on the FMCE. Both CTT and IRT measures of change on the FMCE will be presented, and the assumptions underlying each of these approaches will be discussed and contrasted. The methods section then describes the sample used in this study and the CTT and IRT analysis results (linked to specific items on the FMCE). It will be shown that an examination of how item difficulty changes from pre- to post-test can provide researchers with more detailed information about changes in student understanding as compared to aggregate test scores. By considering the items as an indicator of change, one can isolate and examine specific aspects of Newtonian thinking. Further, there exists a stronger statistical basis for making these item comparisons under the assumptions of IRT (as opposed to CTT), which will be discussed in the methods section. Finally, the discussion section will synthesize the results of the study and suggest future directions for research.

The Force and Motion Conceptual Evaluation

The FMCE was designed to characterize students' conceptual understanding of Newtonian mechanics (Thornton & Sokoloff, 1998). More specifically, it is intended to measure student understanding of kinematics and Newton's laws in one dimension which are generally covered in introductory physics courses. The original purpose of the FMCE was one of formative assessment, as it was intended to be useful as a guide to instruction by indicating in which areas of mechanics student views differed from those of a physicist (Thornton & Sokoloff, 1998). However, many current uses of the FMCE are for characterizing change in students' views, which are more of a summative or evaluative form of assessment rather than a formative one. Thornton, Kuhl, Cummings, & Marx (2009, p. 2) state that "the FMCE was not originally designed to have results analyzed with a single-number score, but to begin our comparison, we felt it necessary to create such a score for the exam." In this way, a students' (and a class') FMCE change scores are used to characterize changing views about physics understanding. These measures of change are then compared across courses in order to make comparisons of instructional efficacy (e.g.,

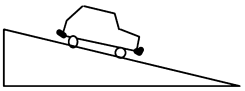
comparing Interactive Engagement [IE][2] courses with traditional courses (e.g., Hake, 1998)). Indeed, the FMCE authors themselves have used the instrument for making these cross-course comparisons. The most common use of this instrument is therefore quite an extension of the original design intention.

The FMCE is used in both algebra and calculus-based general physics courses. These courses usually have large enrollments (between 60 and 500 students per semester). A recent meta-analysis (Ruiz-Primo, Briggs, Iverson, Talbot, & Shepard, 2011) identified 148 comparative studies from 74 papers in physics education, six of which used the FMCE and normalized gain scores to compare an innovative instructional approach to a more traditional one. For example, Cummings, Marx, Thornton, and Kuhl (1999) used the FMCE to evaluate the effectiveness of Interactive Lecture Demonstrations, Cooperative Group Problem Solving, and a standard Studio Physics course. The goal of this work was to characterize the effect of incorporating research-based activities into the Studio Physics course. In another set of studies, Smith and Wittmann (2007) used the FMCE to compare the effect of different tutorials on students' understanding of Newton's Third Law. Overall pre-test FMCE score was used to establish group equivalence. A subset of FMCE items was also used for pre–post comparisons using normalized gain scores.

The FMCE consists of 47 multiple choice items, each with between five and nine answer choices (some of which are purposeful distractors). The authors score the FMCE on a scale of 0 to 33 points, which is based on a composite of the first 43 questions on the instrument.[3] Sets of questions make up categories which are all parts of the construct "Newtonian Thinking." For example, questions 8–10 (see Figure 1) deal with the force on a cart moving on a ramp, questions 11–13 deal with the force on a coin tossed into the air, and questions 27–29 deal with the acceleration of a coin tossed into the air. In order to be deemed a "Newtonian thinker," a student must answer all three of the questions in each of these groups correctly. The composite score derived from the first 43 questions depends upon these categorical groupings. In practical analyses, a composite raw score of about 40% (of the 33 points possible) or below is indicative of non-Newtonian thinking (Thornton et al., 2009).

When the FMCE was created, many physics experts thought the items to be too simple. They "expected that most [students] would answer in a Newtonian way after traditional physics instruction at a selective university." Even after obtaining student responses which showed that

Questions 8-10 refer to a toy car which is given a quick push so that it rolls up an inclined ramp. After it is released, it rolls up, reaches its highest point and rolls back down again. *Friction is so small it can be ignored.*



Use one of the following choices (**A** through **G**) to indicate the **net force** acting on the car for each of the cases described below. Answer choice **J** if you think that none is correct.

(A) Net **constant** force **down** ramp
(B) Net **increasing** force **down** ramp
(C) Net **decreasing** force **down** ramp
(D) Net force zero
(E) Net **constant** force **up** ramp
(F) Net **increasing** force **up** ramp
(G) Net **decreasing** force **up** ramp

_____8. The car is moving up the ramp after it is released.
_____9. The car is at its highest point.
_____10. The car is moving down the ramp.

Figure 1. Common grouping of items dealing with force acting on a car on a ramp.

very few changed their views after traditional instruction, "some professors suggested that perhaps the questions are not significant (or valid or reliable) measures of students' knowledge" (Thornton & Sokoloff, 1998, pp. 338–339). In discussing the validity of the FMCE, Thornton and Sokoloff report quite a difference in student responses between those in IE courses and those in traditional courses. In addition, during development of the FMCE, Thornton and Sokoloff administered the test to "hundreds" of physics faculty, and compared student responses from the multiple choice version to those from an open-ended version which prompts for explanation. They found a very high correlation between these two forms of the test (Saul, 1998). Though there are many pieces of evidence for the validity of the FMCE, there has been no coherent validity argument developed with the depth suggested by frameworks such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Common Approaches to Measuring Change on the FMCE

The authors of the FMCE intended student responses to be analyzed on the basis of raw scores using Classical Test Theory (CTT). In this approach, the composite raw score on the instrument is the statistic for representing the latent variable (in this case, the students' ability to think in Newtonian terms). Under the assumptions of CTT, this observed raw score can be decomposed into a true score (fixed factor) and an error component (random effect). Comparisons between scores (individual students or class

averages) make no distinction in location on the score continuum. For example, a difference of two points near the bottom of the score range is considered to be the same interval as a difference of two points in the middle of the score range. In other words, an interval scale is assumed to exist across the scoring range.

The most common gains analysis applied to the scores from a physics concept inventory (such as the FMCE) is the average normalized gain <*g*>, which Hake (1998) introduced as

$$<g> \equiv \frac{<G>}{<G>_{\max}} = \frac{(Y - X)}{(100 - X)} \qquad (1)$$

where X and Y are the class averages (expressed as a percentage) on a particular physics concept inventory taken at the beginning (i.e., "pre" (X)) and end (i.e., "post" (Y)) of an introductory course in physics. <G> and $<G>_{max}$ represent the class average (raw) gain from pre- to post-tests and the maximum possible class gain respectively (both expressed as percentages). $<G>_{max}$ is the normalizing factor which serves to scale the gains (<G>) and attempts to deal with the observed ceiling effects of the instrument (many of the observed final scores Y are 100%). This normalization makes the assumption of an interval scale across the score continuum. However, in reality, these raw scores are not on an interval scale. They are ordinal at best.

For the data used in this study, <*g*> was calculated to be .64.[4] This is based on the 336 pre/post-test matched raw scores derived from Thornton and Sokoloff's 0–33 point scoring algorithm (see above section on the FMCE). This is a very high value for <*g*> compared to those reported in many other studies, and is nearly in the "high-g" range as defined by Hake (1998). It is important to note that 51 of the 336 matched scores (approximately 15%) had a gain of 1, indicating that they had a perfect score on the post-test.

Commonly, analyses using <*g*> focus on the entire population of students in the course as the unit of analysis. They do not often consider individual students or items as cases for analysis. If such analyses do provide data about responses to particular items or sub-concepts within the framework of Newtonian thinking, it is done under the assumptions of CTT (since the basis for these comparisons is based on raw scores). For example, Thornton and Sokoloff (1998) report the percent correct pre- and post- for various questions on the FMCE as "effect[s] of traditional instruction" (p. 339, Figure 1). Coupled with various correlation studies (involving subgroups by student demographics, education, etc.), measures of <*g*> are accepted by much of the PER community as a basis for making inferences about efficacy of instruction (e.g., Cummings et al., 1999; Meltzer, 2002). For example, based on measures of <*g*> on FCI and FMCE administrations, Cummings et al. (1999) determined that Cooperative Group Problem Solving (an instructional innovation) led to gains in conceptual understanding.

Problems in Measuring Change Using CTT

As mentioned above, raw score measures of change have some limitations. Bereiter (1963) identifies three main "dilemmas": (a) the "over-correction-under-correction dilemma," (b) the "unreliability-invalidity dilemma," and (c) the "physicalism-subjectivism dilemma." I will discuss each of these in turn, as well as ways in which they can be dealt with.

Observed pre-test scores and change scores share the same elements of measurement error (with opposite signs). Consider the following expressions for observed pre-test score (X), observed post-test score (Y), and observed change score ($Y - X$):

$$X = X_t + e_x \qquad (2)$$

$$Y = X_t + G_t + e_y \qquad (3)$$

$$Y - X = G_t + e_y - e_x \qquad (4)$$

In Equations 2–4, X_t represents true pre-test score, e_x represents random error on the pre-test, e_y represents random error on the post-test, and G_t represents the true change score. Note that algebraically, the observed pre-test (X) and change scores (Y-X) share the same error in measurement (e_x) with opposite signs. Because of this, there exists a "spurious negative element" in their correlations. In other words, when raw gain (change) score (Y-X) is regressed on initial score (X), the correlation will likely be understated due to the fact that in part, it is a regression of $-e_x$ on $+e_x$. This shared component of measurement error calls for a correction in the regression of gains on initial scores. This regression itself (of gain score on initial score) is necessary in order to characterize the reliability of the change measurement.

However, the correction for this regression is not straightforward. As Bereiter (1963) notes, the work of Garside shows us that three different methods of solving for this regression (all of which as "plausible") provide us with three widely varying results (an increase in correlation, a decrease in correlation, and an indifference). Depending on which method (or whether some other method, such as a partial correlation) is used, the correction to account for this error sharing element of pre-test and gain scores will either be overstated or understated. Most research reports the uncorrected correlations.

The most common concern with change scores has to do with the "unreliability-invalidity dilemma." Related to the problem with regressing gain scores on initial scores, this dilemma presents itself as a result of these correlations. Researchers would usually like to see a low correlation as a result of this regression, which indicates a higher reliability for the gain scores. However, the problem with this logic is that a very low correlation between gain scores and initial test scores brings into question the validity of the instrument. If these things are not correlated, then it can be argued that the instrument used to obtain the observed pre- and post-test scores do not measure the same construct (i.e., construct definition has changed for the sample from pre- to post-administrations). If the test is therefore not valid, then the change scores on that test lack substantive meaning. This paradoxical relationship has been dealt with in numerous ways, and it can be shown that despite the above logic, gain scores can be reliable without having to show low correlations between gain scores and initial scores (Willett, 1988–89).

For this data set, the correlation between $\langle g \rangle$ and pre-test scores is .026, which is lower than the correlation between raw gain scores and pre-test scores (.174). However, it is still reasonable to believe that this difference is within the range of measurement error in the scores and is therefore subject to Bereiter's first two dilemmas.

The most persistent dilemma in measuring change under the assumptions of CTT has to do with what Bereiter calls "physicalism-subjectivism." This has to do with the scale properties of CTT measurement models, namely that these models assume interval scaling in which equal changes in units anywhere along the scale account for equal changes in the construct being measured. When this dilemma presents itself (as it always does in the measurement of change), Bereiter states that one has "the unpleasant option of sticking with the particular scale units given or some rather arbitrary transformation of them (physicalism), or else abandoning the given units in favor of others that seem to conform to some underlying psychological units (subjectivism)" (1963, p. 5). Although it is easy to pick some transformation of scale and ignore this dilemma, it is especially problematic when many of the raw scores observed are near the extremes. In the present sample, roughly 15% of the students hit the ceiling (i.e., obtained a perfect score) on the FMCE post-test.

Taken together, "these problems seem irresolvable because the change measurements are based on CTT, in which the estimation of item and person parameters is mutually confounded" (Wang & Chyi-In, 2004). The application of IRT to the measurement of change on the FMCE will focus on isolating the items from the persons and examining the change in their difficulties. This is especially appropriate for an analysis of the FMCE, since it can be broken down into subsets of items relating to the construct of Newtonian Thinking as outlined by its designers. CTT-based approaches are not as well suited to this type of analysis.

CTT modeling does not allow the simultaneous assessment of multiple aspects of examinee competence and does not address problems that arise whenever separate parts of a test need to be studied or manipulated. Formally, CTT does not include components that allow interpretation of scores based on subsets of items in the test (Pellegrino, Chudowsky, & Glaser, 2001, pp. 120–1).

Methods

Sample

The current study was conducted at a large public research university in the mountain west, where the FMCE is routinely administered to students in introductory, calculus-based physics courses pre- and post-instruction. The course from which the sample was drawn is the first in a three-course sequence for science and engineering students, is calculus based, and covers a mechanics curriculum. The data for this study come from the spring semester of 2004, which was taught using IE methods. Specifically, the course instructor utilized clickers and the Peer Instruction model (Mazur, 1997), and made use of Learning Assistants (LAs; Otero, Finkelstein, McCray, & Pollock, 2006). The LAs worked primarily in the associated recitation/lab sections (~25 students in each) which used the Washington Tutorials in Introductory Physics (McDermott & Shaffer, 2002). The course also used an online interactive homework system (CAPA: Computer-Assisted Physics Assignments) and a help room for physics students which was staffed daily from 9:00–5:00. It should also be noted that the course instructor was very experienced in teaching using IE methods.

The total number of FMCE pre-test respondents was 468, and total number of post-test respondents was 410. Matched pre- and post-test data exist for 336 students. This is important to note because any gains analysis under CTT can use only these 336 matched student responses. The IRT-based approach to examining change through item difficulty analysis is able to use all respondent data (468 pre- and 410 post-, representing 531 distinct respondents in total). This is because the analyses focus on the "items" themselves, rather than the respondents, and in IRT the estimation of item and person parameters is not mutually confounded, as in CTT.

Approximately, 75% of the respondents were male, and 25% were female. Other background variables such as race and socioeconomic status (SES) were not available.

Analysis

The initial analysis treated all 47 items on the FMCE as independent and dichotomously scored. I used a Rasch model (Bond & Fox, 2007; Rasch, 1980) to estimate the item difficulty parameters for each of the 47 items on the pre-test (n = 468).

$$P(X_{is}=1|\Theta_s,\beta_i)=\frac{\exp(\Theta_s-\beta_i)}{1+\exp(\Theta_s-\beta_i)} \quad (5)$$

This one-parameter logistic model (Equation 5) estimates the probability of correct response to an item i by a person s. In the model, X_{is} represents the response of person s to item i, Θ_s represents the ability estimate (i.e., trait level) of person s, and β_i is the difficulty of item i. For the purposes of this analysis, the person ability estimate (Θ_s) was constrained to have a mean of zero while item difficulty (β_i) was free to be estimated by the modeling software.

I used the IRT modeling software ConQuest (Wu, Adams, & Wilson, 1997) to estimate the item difficulty parameters from the pre-test data. Once the pre-test item difficulties were obtained, a second data set was created that included both pre-test and post-test items (94 items total) for all respondents ($n = 531$). In this fashion, I could command the software to freely estimate the post-test item difficulties while anchoring the pre-test item difficulties which were previously modeled. Direct comparisons could then be made between the values obtained for pre-test and post-test item difficulties for each item.

I also conducted a secondary analysis in which I divided the FMCE into 11 "testlets" (Wainer & Kiely, 1987) based on content groupings similar to those discussed above. This was done in order to deal with the violation of the local independence assumption of IRT. Due to the fact that groupings of items shared common answer pools, it is reasonable to assume that items within these groups were locally dependent on one another. In this analysis, I used a Partial Credit Model (PCM; Masters, 1982) to analyze the resulting polytomous testlet item data. This model is part of the Rasch family of IRT models, and is given by Equation 6.

$$P_{ix}(\theta)=\frac{\exp\left[\sum_{j=0}^{x}(\theta-\delta_{ij})\right]}{\sum_{r=0}^{m_i}\left[\exp\sum_{j=0}^{r}(\theta-\delta_{ij})\right]} \quad (6)$$

This equation gives the probability of a person with ability Θ responding to item i with response x (where item i is scored from $x = 0$ to m_i [the maximum number possible for item i]). The category response is denoted by j, and the parameter δ_{ij} represents step difficulty—the location on the latent ability continuum where a respondent has a 50% probability of a response in category x relative to category x-1.

Again using the ConQuest software (Wu et al., 1997), I ran a PCM on the pre-test response data ($n = 468$) in order to obtain item parameter estimates. These estimates were used to anchor the subsequent run, which included both pre- and post-test responses ($n = 531$). Again, in the first model run, person ability estimates were constrained to have a mean of zero so that item parameters were free to be estimated by the model. The second (combined) run had double the number of testlets (22 total). In this run, pre-test item (testlet) difficulty parameters were anchored to those obtained in the first run, and post-test item (testlet) difficulty parameters were estimated relative to these pre-test values. The resulting item parameter estimates from both runs (pre-test items from the first run and post-test items from the second run) serve as the basis for this secondary analysis.

The description of the IRT models used makes explicit two of the greatest limitations of IRT: (a) estimation procedures for both person ability and item difficulty are complex and not straightforward, and (b) many practitioners and researchers lack the knowledge and experience to carry out such procedures and interpret the results. CTT-based calculations and score interpretations are quite intuitive and well accepted by many educators and researchers. IRT can appear to be a sort of "black box" which does not lend itself to widespread adoption and use.

Results

Dichotomous Rasch Analysis

For both analyses, I express changes in item difficulty from pre- to post-test in terms of effect sizes. Equation 7 gives the effect size (E) calculation based on CTT item difficulty, and Equation 8 gives that for IRT item difficulty estimates.

$$E_{CTT}=\frac{(p_{post}-p_{pre})}{SD_{p_{pre}all\_items}} \quad (7)$$

$$E_{IRT}=\frac{(|\beta_{post}-\beta_{pre}|)}{SD_{\beta_{pre}all\_items}} \quad (8)$$

In Equation 7, p_{post} represents item difficulty for the item in the post-test, p_{pre} represents the item difficulty for the item

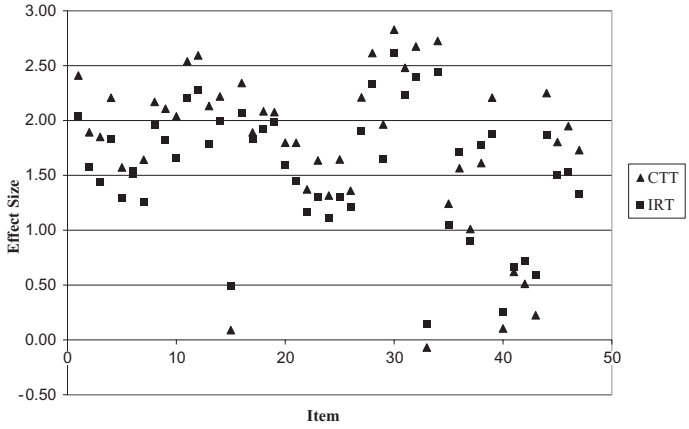


Figure 2. CTT and IRT change in item difficulty effect sizes.

Table 1
Lowest Change in Item Difficulty Effect Sizes for Both CTT and IRT Models

CTT Lowest Item Difficulty Effect Sizes		IRT Lowest Item Difficulty Effect Sizes	
Item	Effect Size	Item	Effect Size
37	1.01	37	.90
41	.62	42	.72
42	.51	41	.67
43	.22	43	.59
40	.10	15	.49
15	.09	40	.26
33	−.07	33	.14

CTT = Classical Test Theory; IRT = Item Response Theory.

in the pre-test, and $SD_{p\text{-}pre\_all\_items}$ is the standard deviation for all 47 pre-test item difficulties (p values). In Equation 8, β_{post} represents IRT item difficulty parameter for the item in the post-test, β_{pre} represents the IRT item difficulty parameter for the item in the pre-test, and $SD_{\beta\text{-}pre\_all\_items}$ is the standard deviation for all 47 pre-test item difficulties (β values). The absolute value is taken for the difference between post- and pre-item difficulties in the IRT calculation due to the fact that the scale is inverted relative to the p values in CTT. In IRT for example, β values become smaller (even negative) as the item gets easier. Plots of CTT and IRT change in item difficulty effect sizes for the first (47 item) analysis are shown in Figure 2. In comparing the two plots visually, one notices a general compression in effect size for change in IRT difficulties relative to that for CTT difficulties.

Note that a direct quantitative comparison between the CTT and IRT item difficulty effect sizes is not possible due to differences in variance and scale. For example, the CTT item difficulties for the pre-test items have an SD = .24, while the IRT item difficulties for the pre-test have an SD = 1.71. Because the CTT difficulties are on an ordinal scale and the IRT difficulties are on an interval scale, normalizing both sets of difficulties for direct comparison is not a statistically sound strategy either. Because of these issues, I will discuss the item difficulties from the two measurement models as they pertain to each item without quantitatively comparing the two different effect size measures.

Discussion of dichotomous Rasch analysis. In each of the plots of change in item difficulty effect size, there are groups of items that clearly have lower effect sizes than the others. Although the order of the lowest effect size items is slightly different in the IRT and CTT models, the items in this grouping are the same (see Table 1) for each measurement model.

Item 37 is one of a group of questions about a car pushing a truck and deals with Newton's Third Law. Fifty-eight percent of respondents answered this question correctly on the pre-test, and 83% answered it correctly on post-test. An interesting question to ask is why the change in item difficulty for this question is so much lower than that for the related questions, 35, 36, and 38? Items 36 and 38 were extremely difficult for pre-test respondents (8% and 7% correct, respectively) and deal with Newton's Third Law and the concept of acceleration in the same situation. Part of the low effect size for item 37 can be accounted for by the fact that it was the easiest of these four items to begin with, and therefore did not have as much room to change. I will further examine this grouping in the polytomous testlet item analysis, as these items together make up one the testlet groupings (testlet 8).

Items 40 through 43 ask the respondent to choose appropriate velocity–time graphs to describe the motion of a car in different situations. These items were fairly easy for pre-test respondents (71–90% correct) and very easy for post-test respondents (86–95%). Again, I will further examine these items as a group (these items comprise testlet 10) in the next section.

Item 15 asks the respondent to choose the force-time graph which represents a car at rest. This item was very easy both on the pre-test and post-test, with 94 and 97% (respectively) answering it correctly. Again, because it was relatively easy on the pre-test, there is not much room for growth or change.

Item 33 was very easy for respondents on both pre- and post-tests. It deals with a collision between vehicles of equal mass and asks respondents about the forces acting on the vehicles during the collision. Similar to the above interpretation, because this item has such a low difficulty, there is no room for change.

Table 2
Highest Change in Item Difficulty Effect Sizes for Both CTT and IRT Models

CTT Highest Item Difficulty Effect Sizes		IRT Highest Item Difficulty Effect Sizes	
Item	Effect Size	Item	Effect Size
30	2.83	30	2.61
34	2.73	34	2.44
32	2.67	32	2.40
28	2.61	28	2.33
12	2.59	12	2.28
11	2.54	31	2.23
31	2.48	11	2.21

CTT = Classical Test Theory; IRT = Item Response Theory.

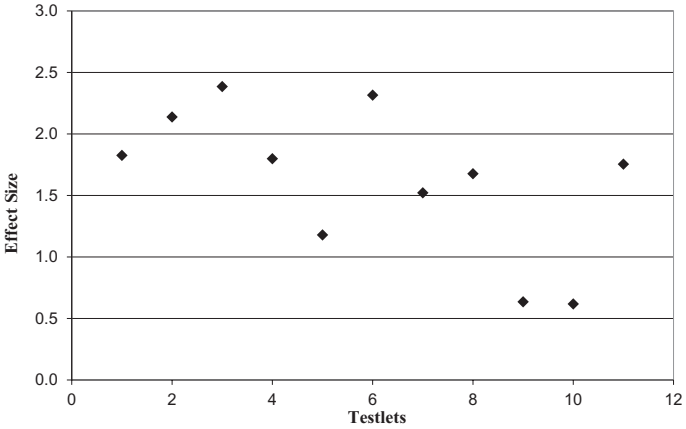


Figure 3. Testlet change in item difficulty effect size.

At the other end of the effect size range, there are similar groupings of items that have the highest change in item difficulty effect sizes under both the CTT and IRT models. These items and their effect sizes are shown in Table 2.

Items 30, 31, 32, and 34 are all part of the same grouping on the FMCE and deal with collisions and Newton's Third Law paired forces. These items were all quite difficult for respondents on the pre-test (between 18 and 28% answered them correctly) and were somewhat easy for respondents on the post-test (between 83 and 87% answered them correctly). These changes indicate a large growth in student understanding regarding this particular topic. Item 33 (which is also part of this grouping) was discussed above and was one of the easiest items overall and therefore did not have much room for change. Together, items 30 through 34 make up testlet 7 and will be discussed in the next section.

Item 28 is one of a group of three items (in testlet 6) which asks students about the acceleration of a coin tossed straight up into the air. This particular item asks about the acceleration at the top of the trajectory. Nineteen percent of students answered this correctly on the pre-test, and 82% answered it correctly on the post-test. The idea that the coin has an acceleration equal to -9.8 m/s^2 (the acceleration due to gravity, g) at the top of its trajectory is a difficult concept for students to understand.

Items 11 and 12 also refer to a coin tossed into the air, but instead of asking students about the acceleration of the coin, these items ask students about the force acting on the coin. For both items, 19% of students responded correctly on the pre-test, and 80–82% responded correctly on the post-test. It is reasonable to think that on the pre-test, the alternative conception was that students assumed that since the coin was either moving upward (question 11) or motionless at the top (question 12), then there could not be a downward force on the coin (the force due to gravity). Because questions 12 and 28 (discussed above) are so closely related, it is not surprising to see low pre-test scores on item 12 after having examined item 28. Data from items 11 and 27 indicate discordant thinking on the part of the respondents regarding force and acceleration on the upward-moving coin.

In the next section, I will examine many of these items grouped together into testlets. Because I do not have set criteria for success on each testlet that can be expressed as p-values (under CTT assumptions), the discussion will deal only with testlet item difficulties as obtained by the polytomous PCM analysis.

Polytomous PCM Analysis

As would be expected, the average item (testlet) difficulties decreased from pre-test to post-test (see Figure 3). What is worth looking at in detail is the relative change in difficulty between testlets and the content of each testlet. For example, the largest magnitude decrease in testlet difficulty is seen in testlets 3 and 6. Each of these testlets decreased in difficulty by an effect size of about 2.3 effect size units. These two testlets each had to do with the same physical phenomenon (the coin tossed into the air). On the other hand, the lowest magnitude decrease in testlet difficulty is seen in testlet 9 (Newton's Third Law) and testlet 10 (one-dimensional motion and velocity-time graphs). Testlet 10 was the easiest item to begin with (it is composed of items 40–43, which are discussed above), so its difficulty could not change much from pre-test to post-test. Testlet 9, however, was of average difficulty initially.

Discussion of polytomous PCM analysis. Given the above information about changes in item difficulties from pre-test to post-test, I am now in a position to make some initial inferences about student learning in the specific areas of Newtonian thinking. The largest gains made by

this class were on the items dealing with the force and acceleration of a coin tossed into the air (testlets 3 and 6). Although the items in these groups were some of the more difficult ones on the pre-test, they were not the most difficult. Testlets 2, 1, and 4 were all more difficult for the students. Therefore, it is reasonable to think that the large changes in item difficulty effect size for testlets 3 and 6 are not merely artifacts of their initial difficulty. It would be interesting to investigate how the topic of free fall was represented in the course curriculum and instruction relative to other topics, such as Newton's Second Law (testlet 1, which showed lower gains).

The areas that showed the lowest change in item difficulty effect size were testlets 5 and 9. Testlet 9 consisted of only one item which deals with Newton's Third Law. Thirty-one percent answered this item correctly on the pre-test, and 84% answered it correctly on the post-test, making it an item of mid-range difficulty initially. Testlet 5 presents the student with acceleration vs. time graphs related to the motion of a car on a ramp. These items were also of mid-range difficulty initially (32 to 51% answered correctly on pre-test) and of moderate difficulty on the post-test (73 to 80% answering correctly). It is somewhat surprising that these items (which represent concepts basic to Newtonian thinking, namely kinematics in one dimension) were not easier on the post-test.

Conclusion

In answering the first research question (At the item level, how do IRT approaches to the measurement of change on the FMCE compare to CTT measures of change?) I find that the changes in item difficulty as determined by the two measurement models are not all that different. In the first analysis, the change in item difficulty effect sizes as found under the CTT and IRT models lead one to examine the same sets of questions. Groupings of items that had the highest and lowest change in item difficulty effect sizes were the same regardless of the measurement model used. Although I did not have a basis for quantitatively comparing the effect size measures from both models, a qualitative comparison shows that they are quite similar, but that the range of IRT effect sizes was compressed relative to those for CTT. Based on these findings, I would recommend that researchers should continue to use CTT measurement models but should consider examining changes in item performance as well as student performance. That said, there is a stronger statistical basis for making claims based on such analyses under the IRT measurement model. The obvious trade-off is ease of interpretability for audiences not familiar with IRT. As stated above, this is a major limitation of using IRT. Lack of familiarity with probabilistic modeling of abilities and item difficulties makes modeling, interpretation, and communication of results difficult.

In addressing the second research question (Do the IRT approaches to measuring change on the FMCE lead us to make different inferences about student learning of Newtonian Physics?), the answer is less clear. Using the dichotomous 47-item Rasch analysis, the answer would be "no." However, using the testlet-based approach and a polytomous Partial Credit Model, I might have a different answer. In order to deal with the violation of the assumption of local independence, items with common answer pools were grouped into testlets. The secondary analysis of the change in item difficulty effect sizes for these testlets provided information that was not available under the CTT measurement model. Specifically, these common item groupings that dealt with similar content could be more easily compared to one another so that different inferences could be made about these groups. The take-home message from this analysis is to consider analyzing conceptually coherent item groupings in addition to aggregate test scores. But a more nuanced implication is that one must have a theory for defining such item groupings, which is a tenet of construct-based measurement and IRT modeling. In the case of these analyses, that theory was both statistical (based on local item dependence) and content oriented. Such groupings need to be theoretically defined in order to support the analyses used and inferences drawn. A set of items that look similar may not constitute a theoretically based grouping upon which inferences can be made. Further, from a validity standpoint one must understand the very real limitations of choosing a subset of items from a test. In doing so, the construct has changed, and previous validity evidence may no longer support such a use.

From a statistical standpoint, the next steps in this line of research should focus attention in two areas: (a) developing a non-parametric method for comparing change in item difficulties under the two measurement models, and (b) using this method to make comparisons of gains (from the perspective of change in item performance) between different semesters of the same course. Once researchers have a sound statistical basis for making these between-semester comparisons, we can better compare gains from the perspective of changing item performance to those characterized by the normalized gain $\langle g \rangle$. From the standpoint of science educator and science education researcher, future research should examine the degree to which the analysis of conceptually coherent (e.g., theoretically defined) item

groupings can provide insight into change in students' conceptual understanding, while acknowledging the potential threats to validity that such an approach might introduce. The current use of aggregate scores from concept inventories may be too blunt an instrument in some cases, especially when we also have at our disposal a much sharper instrument.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison: University of Wisconsin Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: L. Erlbaum.

Bonham, S. W., Deardorff, D. L., & Beichner, R. J. (2003). Comparison of student performance using web and paper-based homework in college-level physics. *Journal of Research in Science Teaching*, *40*(10), 1050–1071.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—Or should we? *Psychological Bulletin*, *74*(1), 68–80.

Cummings, K., Marx, J., Thornton, R. K., & Kuhl, D. (1999). Evaluating innovation in studio physics. *American Journal of Physics*, *67*(S1), S38–S44.

Finkelstein, N. D., & Pollock, S. J. (2005). Replicating and understanding successful innovations: Implementing tutorials in introductory physics. *Physical Review Special Topics—Physics Education Research*, *1*, 010101-1–010101-13.

Hake, R. (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74.

Hake, R. (2002). Lessons from the physics education reform effort. *Conservation Ecology*, *5*(2), art. 28.

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force concept inventory. *Physics Teacher*, *30*(3), 141–158.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mazur, E. (1997). *Peer instruction: A user's manual*. Upper Saddle River, NJ: Prentice Hall.

McDermott, L. C., & Shaffer, P. S. (2002). *Tutorials in introductory physics*. Upper Saddle River, NJ: Prentice Hall.

Meltzer, D. E. (2002). The relationship between mathematics preparation and conceptual learning gains in physics: A possible "hidden variable" in diagnostic pretest scores. *American Journal of Physics*, *70*(12), 1259–1268.

Otero, V., Finkelstein, N., McCray, R., & Pollock, S. (2006). Who is responsible for preparing science teachers? *Science*, *313*(5786), 445–446.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pollock, S. J. (2004). *No Single Cause: Learning Gains, Student Attitudes, and the Impacts of Multiple Effective Reforms*. Paper presented at the Physics Education Research Conference, Sacramento.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press.

Ruiz-Primo, M., Briggs, D. C., Iverson, H. I., Talbot, R. M., & Shepard, L. (2011). Impact of undergraduate science course innovations on learning. *Science*, *331*(6022), 1269–1270.

Saul, J. M. (1998). *Beyond Problem Solving: Evaluating Introductory Physics Courses Through the Hidden Curriculum*. PhD Dissertation. University of Maryland College Park.

Smith, T. I., & Wittmann, M. C. (2007). Comparing three methods for teaching Newton's third law. *Physical Review Special Topics—Physics Education Research*, *3*(2), 020101-1–020101-11.

Thornton, R. K., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing the force and motion conceptual evaluation and the force concept inventory. *Physical Review Special Topics—Physics Education Research*, *5*(1), 010105-1–010105-8.

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation. *American Journal of Physics*, *66*(4), 228–351.

Van Domelen, D. J., & Van Heuvelen, A. (2002). The effects of a concept-construction lab course on FCl performance. *American Journal of Physics*, *70*(7), 779–780.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201.

Wang, W., & Chyi-In, W. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement*, *64*(5), 758–780.

Willett, J. B. (1988–89). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M. L., Adams, R. J., & Wilson, M. (1997). ConQuest Generalised Item Response Modeling Software: Australian Council for Educational Research.

Author's Notes

[1] http://www.compadre.org/psrc/items/detail.cfm?ID=924 and http://listserv.buffalo.edu/cgi-bin/wa?A0=physlrnr-list

[2] Interactive Engagement (IE) is "designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors" (Hake, 1998).

[3] Questions 44–47 deal with mechanical energy and are usually not included in the analyses. It is not always clear if the same scoring strategy is followed in different analyses using the FMCE. For example, Cummings et al. (1999) do not include questions 44–47 and explicitly cite Thornton in their discussion of scoring, who also omits question 6 from some analyses. In short, there appears to be some variability in the way researchers score the FMCE responses.

[4] This value of $\langle g \rangle$ is based on using the class average pre- and post-test scores, as described above in the explanation of the equation for $\langle g \rangle$. Another approach is to average the individual student gains and use this as a measure of the class $\langle g \rangle$. Using this method yields a $\langle g \rangle$ of .66 for the same data.