Supplemental Material

I. ITEM RESPONSE THEORY MODELS AND FIT

Many Item Response Theory (IRT) models have been used to explore the FCI. This work used the 2-parameter logistic (2PL). The 2PL model assumes that each item, j, has a discrimination, a_j, and a difficulty, b_j. The probability, π_{ij}, that participant i answers item j correctly is given by

$$\pi_{ij} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \tag{1}$$

where θ_i models the ability of student i to answer any item correctly. Some authors rescale a_j to map the logistic function approximately onto the normal-ogive function. We report the untransformed discrimination a_j; that is, we work in the logistic metric as opposed to the normal metric.

Multiple studies have applied IRT to the FCI using models related to the 2PL model. Two studies applied the Rasch model [1, 2], a simplification of the 2PL which sets the discrimination to one ($a_j = 1$). An extension to the 2PL model that allows multiple ability traits for each student, multi-dimensional IRT, was used as an alternative to factor analysis for the FCI [3]. Morris *et al.* applied an alternate method called "Item Response Curve" analysis which replaces the ability, θ, estimated by IRT with the overall FCI score [4].

The 3PL model extends the 2PL model with a third parameter, c_{ij}, designed to model random guessing behavior. Wang and Bao [5] used the 3PL model to investigate the FCI. We tested the 3PL model on Sample 1. It improved model fit somewhat, but the guessing parameters extracted were much too large to be credible for posttest results in a course producing the strong conceptual performance of Sample 1. Similar departures from model fit can be seen in the 3PL plots of Wang and Bao [5], where many curves diverge from the data at low ability. Item Response Curve analysis was compared with Wang and Bao's IRT analysis [6] and found very small guessing parameters, indicating the 2PL model may be more appropriate. Additionally, uniform guessing parameter models have been challenged for distractor-driven tests [7]. For these reasons, this study will employ the 2PL model, which is also the most closely related to CTT.

IRT introduces a model for student responses (Eqn. 1); the degree to which this model fits the data was investigated. IRT model fit can be evaluated for each item by dividing the students into G groups by their estimated ability, θ, and then estimating the goodness of fit between the predicted mean of the group given by the 2PL model and the observed mean [5, 8–10]. This produces a χ^2 distributed statistic with $df = G-2$ degrees of freedom, because the 2PL model estimates two parameters per item. Chi-squared tests have known problems with rejecting the null hypothesis of good item fit for large samples [9]. To overcome this limitation, Cramer's V effect size statistic was calculated, $V = \sqrt{\chi^2/(df \cdot N)}$, for both male and female students [11] and is reported in Table III in the main text. For V, $V = 0.1$ represents a small effect, $V = 0.3$ a medium effect, and $V = 0.5$ a large effect. The number of groups used varies by study; we selected a G that ensured at least 100 students were in each group, leading to $G = 10$ for women and $G = 33$ for men. While some items were detected as significantly not fitting the 2PL model, no misfit represented even a small effect size. Detection of some misfitting items was expected because of the large sample size of the study. As such, the 2PL parameter estimates should be accurate for this dataset. Item Characteristic Curves were examined for all items. The plots for both male students for all items and female students for most items had similar visual fit to those presented in Wang and Bao [5]. For all curves, significant misfit was a result of variance between nearby bins and not an overall failure of the 2PL model to fit the data.

For IRT analysis using the 2PL model, estimates of minimum required sample size vary, with some authors suggesting that a minimum of 200 is acceptable while others that samples of 500 are required [12]. While Sample 1 has sufficient male and female students, there were too few female students in Samples 2 and 3 for accurate parameter estimation.

II. DIFFERENTIAL ITEM FUNCTIONING STATISTICS

A substantial number of DIF statistics have been investigated for IRT; we report Lord's statistic, L, which compares the difference in difficulty parameters for the Rasch model with the average difference in difficulty [13]. Multiplying by 2.35 projects Lord's statistic onto the ETS Delta scale [14, 15],

$$L_i = 2.35 \cdot \left(b_i^F - b_i^M - \frac{1}{n}\sum_{j=1}^{n}(b_j^F - b_j^M) \right), \tag{2}$$

where b_i^F is the female difficulty on item i, b_i^M the male difficulty, and $n = 30$, the number of items.

Lord's statistic was selected both because it corresponds to an effect size measure on the Delta scale and because it allows comparison with Osborn Popp, Meltzer, and Megowan-Romanowicz's large study of DIF in high school students [1]. They used IRT with the Rasch model and a DIF statistic computed as the difference in the b difficulty parameter between men and women [13]. Their population had an average difference in difficulty parameters of approximately zero, and as such, their difference statistic was equivalent to Lord's statistic, L, before multiplying by 2.35.

Lord's statistic is calculated with the difficulty parameters in the Rasch model in which the discrimination is set to one ($a_i = 1$). The difficulty parameters reported in Table III in the main text are for the 2PL model; therefore, Eqn. 2 above cannot be used to compute Lord's statistic using b_i values from Table III. Further, the difficulty parameter calculated in the Rasch model for item 29 was reasonable, allowing L to be calculated for this item even through the difficulty and discrimination in the 2PL model were problematic.

DIF analysis with the MH statistic groups students into strata. The finest grain possible divides the students into groups with the same total test score; less fine-grained strata can be formed by dividing students into ranges of test scores. For example, Dietz *et al.* divided students into five quantiles [16]. The large number of participants in Sample 1 allowed the division by test score. For Sample 1, both stratifications were compared and while $\Delta\alpha_{MH}$ was somewhat different between the methods, both yielded the same classification of DIF effect size on the ETS Delta scale.

III. PRETEST RESULTS

Fig. 1 shows the FCI pretest results for Sample 1. The five substantially unfair questions identified in the posttest (14, 21, 22, 23, 27, Fig. 1 in main text) were among the most unfair questions in the pretest plots. However, many additional questions were also substantially more difficult for women. The IRT variance for women was also substantially higher than in the posttest. Many pretest differences were reduced by instruction and many questions moved substantially closer to the fairness line in the posttest, except items 14, 21, 22, 23, and 27.

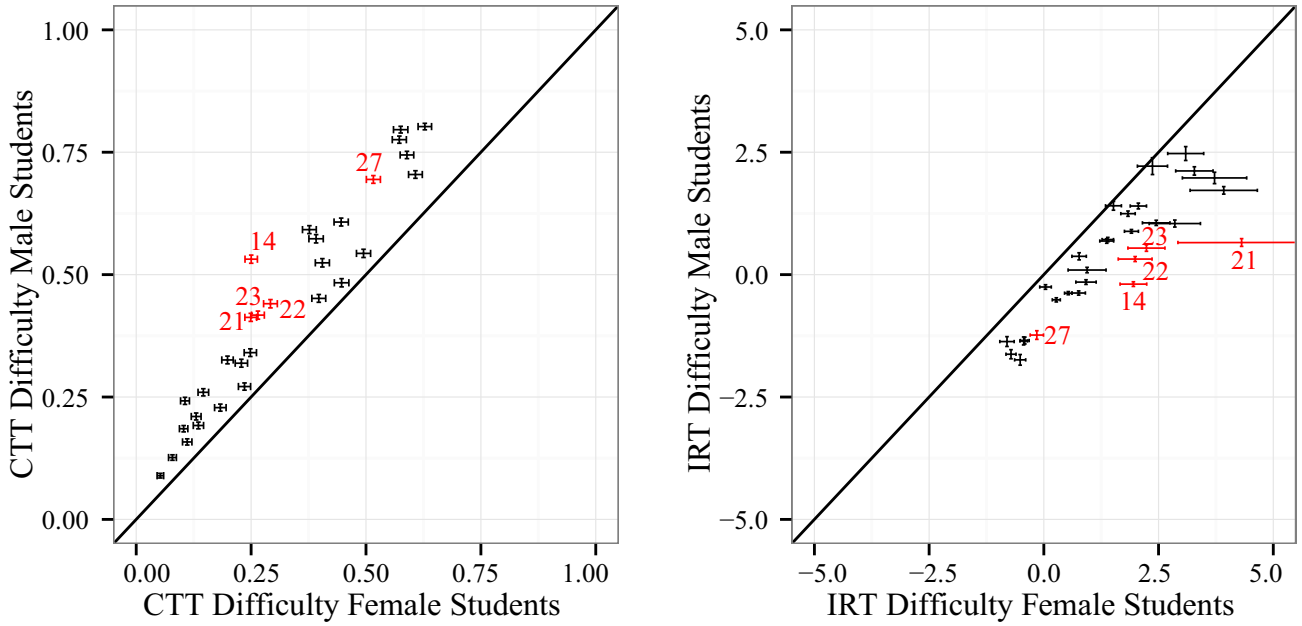

FIG. 1. CTT and IRT pretest results for Sample 1. Items 14, 21, 22, 23, and 27 are marked in red and labeled. A line of slope one is drawn to allow comparison of male and female difficulty. Error bars represent one standard deviation in each direction.

DIF analysis was also performed on the Sample 1 pretest. With the much larger variance seen in Fig. 1 and the generally weaker pretest performance of women, few items were detected as significantly biased. The DIF results for the FCI pretest for Sample 1 detected only item 14 as having large DIF; items 4, 12, 19, and 26 demonstrated small to moderate DIF. This difference between pretest and posttest is consistent with the observation that women close the score gap with men on many problems post instruction. Because DIF stratifies by overall test score, a smaller gap can be considered unfair on the posttest than the pretest if the overall posttest gap is smaller than the pretest gap.

This study found a 20-item unbiased version of the FCI. The pretest gender gaps changed little on the reduced unbiased instruments. For Sample 1, the gender gap on the 20-item FCI was 9.9% which was somewhat smaller than the gender gap of 11.9% on the original 30-item FCI. Further removing item 29 increased the gap to 10.1%. For Sample 2, the gender gap on the 20-item instrument was 10.3% which was somewhat smaller than the gender gap of 12.3% on the original 30-item FCI. Further removing item 29 increased the gap to 10.6%.

IV. RELIABILITY AND CORRELATION ANALYSIS

Cronbach's alpha provides a measure of the overall reliability of an instrument. If alpha increases with the removal of an individual item, that item detracts from the overall instrumental reliability and should be a candidate for elimination. Only posttest results were explored for this analysis. For Sample 1, the FCI was reliable with $\alpha = 0.84$ overall, male students $\alpha = 0.84$, and female students $\alpha = 0.83$. For male students, dropping item 29 increased alpha, while there was no item that could be removed to increase alpha for female students. For Sample 2, overall $\alpha = 0.90$ with $\alpha = 0.91$ for men and $\alpha = 0.81$ for women. For male and female students, there was no item whose removal increased alpha. For Sample 3, overall $\alpha = 0.86$: with $\alpha = 0.85$ for male students and $\alpha = 0.82$ for female students. Removing item 15 increased the overall alpha for both male and female students. These reliability values were consistent with those reported in Lasry *et al.* [17] and show that the FCI has strong internal consistency across a variety of instructional settings. Cronbach's alpha of 0.7 is considered acceptable reliability; alpha of 0.9 is required for higher stakes tests [18].

To further investigate reliability, the correlation coefficient between items can be calculated. In general, if a student answers one item on a test correctly, the probability of answering a second item correctly should increase; item scores should be positively correlated. Jorion *et al.* [19] calculated tetrachoric correlations which assume the dichotomous variable, whether the question was correct or incorrect, was derived from an underlying normal continuum. This assumption seems unnatural for multiple-choice physics questions where the student must either answer completely correctly or incorrectly. Instead, we report the Pearson correlation, which for two dichotomous variables is the ϕ coefficient [20]. Tetrachoric correlations were also calculated and in all cases had absolute values greater than $|\phi|$. The significantly negatively correlated ($p < 0.05$) item pairs in Sample 1 were: male students, $\{23, 29\}$ and $\{29, 30\}$ and female students $\{8, 21\}$, $\{15, 27\}$, and $\{29, 30\}$. In Sample 2, there were no significantly negatively correlated item pairs for male students; for female students, only items $\{12, 29\}$ were significantly negatively correlated. For Sample 3, no question pairs were negatively correlated for men, while $\{7, 15\}$ and $\{9, 12\}$ were significantly negatively correlated for women. Both the correlation analysis and Cronbach's alpha support the identification of item 29 as problematic. Many of the items which were negatively correlated are identified as unfair in subsequent DIF analysis: items 9, 12, 15, 21, 23, and 27.

The ϕ coefficient above is mathematically similar to the ϕ coefficient in Table III in the main text; however, their use is conceptually different. Above, ϕ is used as a measure of association, so large ϕ indicates strongly correlated items. In Table III, ϕ is used as a measure of independence and large ϕ indicates that the item difficulty is different for men and women (small ϕ indicates the difficulty is independent of gender).

[1] S. Osborn Popp, D. Meltzer, and M.C. Megowan-Romanowicz, "Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics," in *2011 American Educational Research Association Conference* (American Education Research Association, Washington, DC, 2011).

[2] M. Planinic, L. Ivanjek, and A. Susac, "Rasch model based analysis of the Force Concept Inventory," Phys. Rev. Phys. Educ. Res. **6**, 010103 (2010).

[3] T.F. Scott and D. Schumayer, "Students' proficiency scores within multitrait item response theory," Phys. Rev. Phys. Educ. Res. **11**, 020134 (2015).

[4] G.A. Morris, L. Branum-Martin, N. Harshman, S.D. Baker, E. Mazur, S. Dutta, T. Mzoughi, and V. McCauley, "Testing the test: Item response curves and test quality," Am. J. Phys. **74**, 449–453 (2006).

[5] J. Wang and L. Bao, "Analyzing Force Concept Inventory with Item Response Theory," Am. J. Phys. **78**, 1064–1070 (2010).

[6] G.A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S.D. Baker, "An item response curves analysis of the Force Concept Inventory," Am. J. Phys. **80**, 825–831 (2012).

[7] C. DeMars, *Item Response Theory* (Oxford University Press, New York, NY, 2010).

[8] R.D. Bock, "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," Psychometrika **37**, 29–51 (1972).

[9] S.P. Reise, "A comparison of item- and person-fit methods of assessing model-data fit in IRT," Appl. Psych. Meas. **14**, 127–137 (1990).
[10] W.M. Yen, "Using simulation results to choose a latent trait model," Appl. Psych. Meas. **5**, 245–262 (1981).
[11] W.J. Conover, *Practical Nonparametric Statistics, Third Edition* (John Wiley & Sons, New York, NY, 1999).
[12] J. Morizot, A.T. Ainsworth, and S.P. Reise, "Toward modern psychometrics," in *Handbook of Research Methods in Personality Psychology*, edited by Richard W. Robins, R. Chris Fraley, and Robert F. Krueger (Guilford Press, Robins, New York, 2009) pp. 407–423.
[13] F. Lord, *Applications of Item Response Theory to Practical Testing Problems* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1980).
[14] P.W. Holland and D.T. Thayer, "An alternate definition of the ETS delta scale of item difficulty," ETS Research Report Series **Research Report RR-85-43** (1985).
[15] R.D. Penfield and G. Camilli, "Differential item functioning and item bias," in *Handbook of Statistics. Vol. 26. Psychometrics*, edited by C.R. Rao and S. Sinharay (Elsevier, Amsterdam, 2007) pp. 125–168.
[16] R.D. Dietz, R.H. Pearson, M.R. Semak, and C.W. Willis, "Gender bias in the Force Concept Inventory?" in *2011 Physics Education Research Conference Proceedings*, Vol. 1413, edited by N.S. Rebello, P.V. Engelhardt, and C. Singh (AIP Publishing, New York, 2012) pp. 171–174.
[17] N. Lasry, S. Rosenfield, H. Dedic, A. Dahan, and O. Reshef, "The puzzling reliability of the Force Concept Inventory," Am. J. Phys. **79**, 909–912 (2011).
[18] J.C. Nunnally and I.H. Bernstein, *Psychometric Theory, Third Edition* (McGraw-Hill, New York, NY, 1994).
[19] N. Jorion, B.D. Gane, K. James, L. Schroeder, L.V. DiBello, and J.W. Pellegrino, "An analytic framework for evaluating the validity of concept inventory claims," J. Eng. Educ. **104**, 454–496 (2015).
[20] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, New York, 1986).