Motivation for the Physics Assessment Evidence Project

Over the last 25 years, student learning in introductory physics classes has often been assessed using legacy instruments such as the Force Concept Inventory (FCI) [3] and the Force and Motion Conceptual Evaluation (FMCE) [4,5] – instruments designed to measure students' conceptual understanding of Newtonian mechanics. These instruments have been critical to the development of Physics Education Research (PER) as a discipline and the recognition that reformed instruction is crucial for the development of student understanding [6]. However, these instruments have serious flaws.

1. **The legacy instruments have substantial psychometric problems which may limit their usefulness for both research and instructional applications.** Since the FCI's introduction, the validity of the instrument has been challenged [7-10]. A substantial strand of research has shown the instrument does not have the factor structure suggested by the authors and that Exploratory Factor Analysis extracts factors for which there is little theoretical support [11,12]. A well-constructed instrument of the length of the FCI or the FMCE should measure some overall construct as well as several subdimensions (sub-constructs) of that construct. These subdimensions should be measured by subscales (groups of items measuring the same sub-construct) within the instrument in order to provide a user of the instrument with a more detailed picture of the construct than is provided by the overall instrument score. Because neither the FCI nor the FMCE were constructed to contain a reliable set of subscales, both instruments only provide an overall score and do not provide the additional detail of subscale scores which would allow an instructor to pinpoint places where instruction needed improvement. Recently, the lack of subscale structure has been tied to flaws within the instrument resulting from the practice of collecting items into item blocks each referring to a common stem (causing unintended correlations between items) and the use of a small subset of isomorphic items with common solution structure [13]. More general psychometric flaws in both the FCI [13-16] and FMCE [17,18] have also been reported with many items having difficulty or discrimination values that would lead them to be flagged as problematic in Classical Test Theory [19]. More recent network analytic studies [20,21] have also suggested some items are not functioning as intended. In general, these issues have led to calls to provide alternate scoring for the FCI [22] and led the authors of the FMCE to introduce an alternate scoring scheme which eliminated multiple items [5].
2. **These instruments have serious and potentially harmful demographic biases.** For example, using a standard method to define an item as fair to a group of students if students in the group with the same overall facility with the material as a reference group score equally to the reference group, a recent large study at three institutions by this project's PI (Henderson) and Co-PI (J. Stewart) and collaborators found five items within the FCI to be substantially unfair to women [16]. These items had been sporadically reported as unfair in research for over 15 years [16, 23-27]. In addition to gender, recent research has shown differences in inventory scores for underrepresented minority (URM) students [28, 29], first-generation college (FGC) students and students from rural areas [30]. In general, these inequities can "reinforce with students the false notion that [they] do not belong in higher education" [31] and more specifically, in physics. Over the last 30 years, the teaching of physics has evolved yet the formative assessments that instructors use to evaluate students' conceptual understanding of Newtonian mechanics within the classroom have remained the same. Therefore, it is crucial that the field of PER develop new and more equitable assessment tools.

3. **Both instruments feature items with artwork and contexts from a different era and represent limited diversity.** They predominantly use white male subjects, feature contexts that may be less familiar to some populations (ice hockey, sleds on icy ponds, objects dropped from planes), and situations that may now be generally unfamiliar (the space shuttle). As an example, an item common to both the FCI and FMCE is shown in Figure 1 (it appears in the FMCE with similar artwork but slightly different text). Not only may the anachronistic nature of the instruments discourage or alienate students, but we note also, for example, that the situation in Figure 1 is offensive to students in some cultures where touching one another with the feet is, at best, improper.

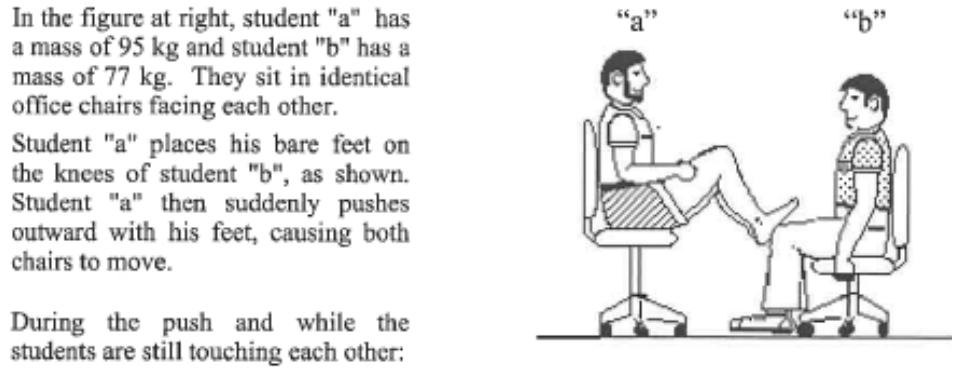

Figure 1: FCI Item 28. An item with a problematic representation and context [3].

The flaws identified in the legacy instruments have come to represent a serious impediment to research in PER involving student understanding. The poor psychometric properties have led some studies to identify structure arising from misleading artifacts of the instrument as real features of student understanding. Many physics education researchers, including the PIs of this project, have explored student understanding of Newtonian mechanics (as measured by the FCI or the FMCE) between different demographic groups. Much of the literature revolves around the well-established "gender gap" where male students outperform female students by 12% on the FCI and the FMCE [32]. The results of these studies must be reexamined considering the fairness issues identified in the FCI. While many reasons have been investigated to try to explain these differences such as prior academic preparation [33-36], cognitive differences [37-40], and psychocultural factors [41-45], there is little agreement within the community as to why there are consistent differences in FCI scores between men and women. In addition to gender, differences in FCI scores between URM and non-URM students have been investigated [28, 29] as well as differences between FGC students and non-FGC students and rural and urban students [30].

Preliminary Findings

As a precursor to this proposed work and as a beginning of the ECD process, the PIs conducted surveys and interviews with 13 introductory physics instructors from nine different institutions across the country including six Predominantly White Institutions (PWI), two Hispanic Serving Institutions (HSI), and one Minority Serving Institution (MSI). **Generally, we learned that the community would benefit from an 'improved' version of the existing tools rather than something completely new.** The instructors mentioned that the existing assessment tools are the best available options for their classes as they work "well enough" to capture the big picture of the students' understanding of the concepts. However, due to the natural limitations of the instruments, it is impossible to see the students' reasoning involved and the problem-solving process. They alluded to having a set of tools that allowed for some flexibility in their assessment strategy. The majority of the instructors in our sample addressed their limited bandwidth and available resources to evaluate or conduct research on the existing assessment tools, but all are open to considering improved versions of the instruments.

References

[1] Mislevy, R. J., Haertel, G., Riconscente, M., Rutstein, D. W., & Ziker, C. (2017). Evidence-centered assessment design. In *Assessing model-based reasoning using evidence-centered design* (pp. 19-24). Springer, Cham.

[2] Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*(4), 454-496.

[3] Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher*, *30*(3), 141-158.

[4] Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *american Journal of Physics*, *66*(4), 338-352.

[5] 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.

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

[7] Lasry, N., Rosenfield, S., Dedic, H., Dahan, A., & Reshef, O. (2011). The puzzling reliability of the Force Concept Inventory. *American Journal of Physics*, *79*(9), 909-912.

[8] Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure?. *The physics teacher*, *33*(3), 138-143.

[9] Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *The physics teacher*, *33*(8), 502-502.

[10] Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The physics teacher*, *33*(8), 503-503.

[11] Scott, T. F., Schumayer, D., & Gray, A. R. (2012). Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics-Physics Education Research*, *8*(2), 020105.

[12] Scott, T. F., & Schumayer, D. (2015). Students' proficiency scores within multitrait item response theory. *Physical Review Special Topics-Physics Education Research*, *11*(2), 020134.

[13] Stewart, J., Zabriskie, C., DeVore, S., & Stewart, G. (2018). Multidimensional item response theory and the Force Concept Inventory. *Physical Review Physics Education Research*, *14*(1), 010137.

[14] Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064-1070.

[15] Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825-831.

[16] Traxler, A., Henderson, R., Stewart, J., Stewart, G., Papak, A., & Lindell, R. (2018). Gender fairness within the force concept inventory. *Physical Review Physics Education Research*, *14*(1), 010103.

[17] Henderson, R., Miller, P., Stewart, J., Traxler, A., and Lindell, R. (2018). Item-level gender fairness in the Force and Motion Conceptual Evaluation and the Conceptual Survey of Electricity and Magnetism. Physical Review Physics Education Research, 14, 020103.

[18] Yang, J., Zabriskie, C., & Stewart, J. (2019). Multidimensional item response theory and the force and motion conceptual evaluation. *Physical Review Physics Education Research*, *15*(2), 020141.

[19] Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

[20] Wells, J., Henderson, R., Stewart, J., Stewart, G., Yang, J., & Traxler, A. (2019). Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis. Physical Review Physics Education Research, 15(2), 020122.
[21] Yang, J., Wells, J., Henderson, R., Christman, E., Stewart, G., & Stewart, J. (2020). Extending modified module analysis to include correct responses: Analysis of the Force Concept Inventory. *Physical Review Physics Education Research*, *16*(1), 010124.
[22] Hudson, R. C., & Munley, F. (1996). Re-score the force concept inventory!. *Phys. Teach.*, *34*(5), 261-261.
[23] Osborn Popp, S., Meltzer, D., and Megowan-Romanowicz, M.C. (2011). Is the Force Concept Inventory biased? Investigating differential item functioning on a test of conceptual learning in physics, American Educational Research Association Conference (American Education Research Association, Washington, DC).
[24] Dietz, R.D., Pearson, R.H., Semak, M.R., and Willis, C.W. (2012). Gender bias in the Force Concept Inventory? in AIP Conf. Proc. 1413, 171.
[25] McCullough, L. and Meltzer, D.E. (2001). Differences in male/female response patterns on alternative-format versions of FCI items, Physics Education Research Conference Proceedings, edited by K. Cummings, S. Franklin, and J. Marx (AIP Publishing, New York), pp. 103–106.
[26] McCullough, L. (2004). Gender, context, and physics assessment, J. Int. Women's St. 5, 20. DIF and Validation ETS Standards for Quality and Fairness, https://www.ets.org/s/about/pdf/standards.pdf, accessed 11/11/2017.
[27] Zieky, M. (2006). Fairness review in assessment, in Handbook of Test Development, edited by S. M. Downing and T. M. Haladyna (Lawrence Erlbaum, Hillsdale, NJ), pp. 359–376.
[28] Henderson, R. and Stewart, J. (2018). Racial and ethnic bias in the Force Concept Inventory. PERC Proceedings [Cincinnati, OH, July 26-27, 2017], edited by L. Ding, A. Traxler, and Y. Cao.
[29] Salehi, S., Burkholder, E., Lepage, G. P., Pollock, S., & Wieman, C. (2019). Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Physical Review Physics Education Research*, *15*(2), 020114.
[30] Henderson, R., Zabriskie, C., and Stewart, J. (2019). Rural and first-generation performance differences on the Force and Motion Conceptual Evaluation. 2018 PERC Proceedings [Washington, DC, August 1-2, 2018], edited by A. Traxler, Y. Cao, and S. Wolf.
[31] Montenegro, E., & Jankowski, N. A. (2017). Equity and assessment: Moving towards culturally responsive assessment. *Occasional Paper*, *29*.
[32] Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?. *Physical Review Special Topics-Physics Education Research*, *9*(2), 020121.
[33] Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological bulletin*, *140*(4), 1174.
[34] Cole, N. S. (1997). The ETS Gender Study: How Females and Males Perform in Educational Settings.
[35] Kost, L.E., Pollock, S.J., and Finkelstein, N.D. (2009). Unpacking gender differences in students' perceived experiences in introductory physics, AIP Conf. Proc. 1179, 177
[36] Miyake, A., Kost-Smith, L.E., Finkelstein, N.D., Pollock, S.J., Cohen, G.L., and Ito, T.A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science 330, 1234
[37] Maeda, Y. and Yoon, S.Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT: R), Educ. Psychol. Rev. 25, 69.
[38] Halpern, D.F. (2012). Sex Differences in Cognitive Abilities, 4th ed. (Psychology Press, Francis & Tayler Group, New York, NY).
[39] Hyde, J.S. and Linn, M.C. (1988). Gender differences in verbal ability: A meta-analysis, Psychol. Bull. 104, 53.

[40] Hyde, J.S., Fennema, E., and Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis, Psychol. Bull. 107, 139.