To θ or not to θ: A simulation study on the validity of IRT
Frequently Asked Questions (FAQs)

- **How can I contact the primary investigator (PI)?**
  - E-mail: JPark@psu.edu
  - Web: IAmJonathanPark.com
- **Where can I find references for the talk?**
  - At the end of this document organized in APA format.
- **Where can I access the simulation code?**
- On my GitHub page: HERE or at this URL: https://github.com/JPark93/Validity-of-IRT
- **What other conditions were simulated?**
  - True person-scores were always normally distributed; $X \sim N(0, 1)$
  - Other distributions of item difficulty, however, were simulated with negatively skewed, and random normal distributions tested as well. Currently, these data have not been thoroughly analyzed
- **What other results were gathered?**
  - Spearman Rank Order correlations were collected to assess the degree to which rank was preserved for each participant
  - Raw rank order proportions were collected to assess the proportion of each simulated sample that was exactly placed by each test methodology
- **What about item parameter recovery?**
  - Item parameter recovery was not assessed for the current simulation study. The 2PLM simulation condition used many of the default settings from the *TAM* package in *R*. Specifically, the simulation utilized the *tam.mml.2pl*() function which utilizes—as its name would suggest—marginal maximum likelihood

(MML). Thus, the rudimentary algorithm built for this simulation is likely to perform comparably to other studies that have relied on MML for estimation and can easily be found online. However, if one is interested in assessing the parameter recovery, I did collect all the pertinent information for one to do so (e.g., true item parameters, item parameter estimates and standard errors).

- **What is the likely effect that human error may have on the results of this simulation?**
    - With a relatively simple Item Response Model (IRM) such as the 2PL, it is likely that the results of the simulation match what one would expect to find from a real-world application. However, it should be noted that this assumes a strict adherence to heuristics and cut-offs. A human researcher may prioritize certain test items due to their theoretical relevance over explicit item functioning. This gap between human-applied and simulated research is likely to widen with the application of polytomous IRMs as far more parameters are being estimated and thus the likelihood of error increases
- **What about polytomous models?**
    - Polytomous models estimate far more parameters than the dichotomous family of IRMs. This significantly increases computational time and was not the current focus of the current exploration. However, future research should look into the effects that poorly functioning items have on the polytomous family of models and whether the increased specificity of such models provide additional insight/benefits over traditional methods

- **Why was item information prioritized in the item selection process? Why not item discrimination as in previous studies?**
    - Within the 2PLM, item information and item discrimination are directly linked together. The item information function for 2PLMs is as follows:

      $$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta) \qquad (1)$$

      where a_i is the discrimination parameter for item i;

      θ is an individual's level of ability on the latent trait being measured;

      $$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

      $$Q_i(\theta) = 1 - P_i(\theta)$$

    - After item parameters are estimated, item information is calculated using a series of quadrature points that span the expected limits of the latent construct. When an individual's ability (θ) and an item's difficulty (b) coincide:

      $$P_i(\theta) = 0.50$$

      $$Q_i(\theta) = 0.50$$

      As $P_i(\theta)$ and $Q_i(\theta)$ sum to 1.00, 0.25 is the largest product of the two. Thus, the expected item information, when there is no discrepancy between an individual's ability and an item's difficulty (i.e., the peak of the item information function) is:

      $$I_i(\theta) = 0.25a_i^2$$

- **How were false positive/negative values calculated?**
    - False positive/negative values were initially saved as 1 or 0 values depending on whether an initially generated ‘true’ score was above/below the 75th percentile. Following that, all other models (e.g., unweighted sums, α-optimization, factor analysis, and IRT) were estimated. The 75th percentile was then calculated for all estimated scores and compared and scored in a similar way to the true scores and subtracted from one another. The results were then totaled and reported as a frequency. Below is a table showing how the false positives and negatives would be calculated:

True	Estimate	Outcome	Classification
0	0	0	Hit
1	1	0	Hit
0	1	-1	False Positive
1	0	1	False Negative

- **Why was the 2PLM selected for this series of simulations?**
    - The 2PLM is perfectly suited for a large body of psychological research by allowing for independent estimation for individual item discrimination and difficulty parameters. Furthermore, the code that was written is readily generalizable to the 3PLM and the 1PLM.

References

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker..

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment*, *20*(1), 55.

Ferrando, P. J., & Chico, E. (2007). The external validity of scores based on the twoparameter logistic model: Some comparisons between IRT and CTT. *Psicológica*, *28*(2).

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of personality and social psychology*, *78*(2), 350.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and psychological measurement*, *62*(6), 921-943.

Mead, A. D., & Meade, A. W. (2010). Test construction using CTT and IRT with unrepresentative samples. In *annual meeting of the Society for Industrial and Organizational Psychology in Atlanta, GA* (Vol. 56).

Mislevy, J. (2008). Using item response theory in survey research: Accommodating complex sampling designs for respondents and items.

Reeve, B. B., & Mâsse, L. C. (2004). Item response theory modeling for questionnaire evaluation. *Methods for testing and evaluating survey questionnaires*, 247-273.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453-468.