Item Characteristic Curves generated from common CTT Item Statistics

Diego Figueiras[1] **John Kulas**[1]

[1] *Montclair State University*

Introduction

Item characteristic curves are frequently referenced by psychometricians as visual indicators of important attributes of assessment items - most frequently *difficulty* and *discrimination*. Assessment specialists who examine ICC's usually do so from within the psychometric framework of either Item Response Theory (IRT) or Rasch modeling. These frameworks provide the parameters necessary to plot the ogive functions. If the curve transitions from low to high likelihood of correct response at a location toward the lower end of the trait (e.g., "left" on the plotting surface), this indicates that it is relatively easy to answer the item correctly. If the curve is sharp (e.g., strongly vertical), this indicates high discrimination; if it is flatter, that is an indication of poorer discrimination - see Figure 1.

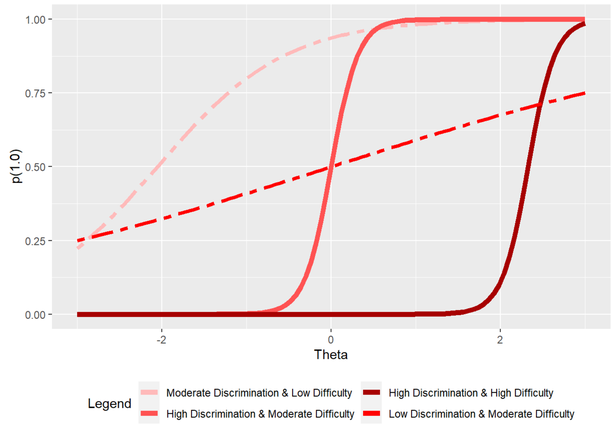


Figure 1: Item characteristic curves reflecting visual differences in difficulty and discrimination.

From a Classical Test Theory (CTT) orientation, item difficulty is most commonly represented by the percent of individuals answering the item correctly (also referred to as a *p-value*). Item discrimination can be conveyed via a few different CTT indices, but the most commonly calculated and consulted index is the corrected item-total correlation.

Method

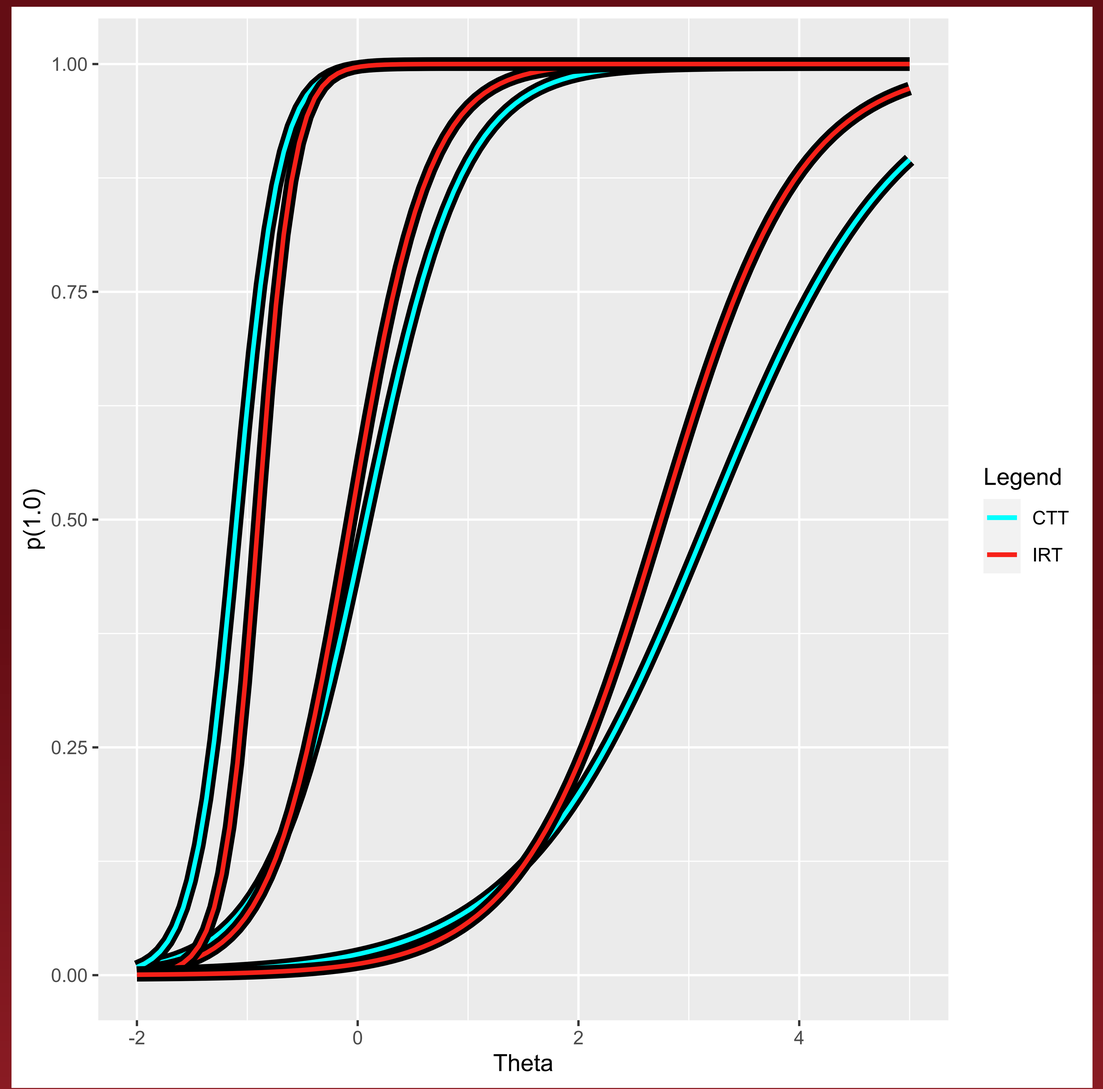
We simulated 10,000 observations across 100 binary items via WinGen **(Han, 2007)**. Because we wanted a range of universally **positive** item discrimination values to model, we specified a mean *a*-parameter value of 2 (*sd* = 0.8). The mean *b*-parameter value was set at 0 (*sd* = 0.5). Next, the `mirt` package **(Chalmers, 2021)** was used to "re"-estimate the IRT via 2PL specification. As for the CTT-derived estimates, **Lord (1980)** provides a conceptual relationship between the IRT parameter and corrected item-total **biserial** correlation:

$$a_i \cong \frac{r_i}{\sqrt{1 - r_i^2}}$$

Kulas et al. (2017), via simulation, provide a less elegant but residual minimizing amendment that: 1) utilizes the more common contemporary corrected item-total **point-biserial** correlation, and 2) captures the influence of item difficulty via specification of a *p-value* derivitive denoted as z_g:

$$\hat{a}_i \cong [(.51 + .02z_g + .3z_g^2)r] + [(.57 - .009z_g + .19z_g^2)\frac{e^r - e^{-r}}{e - e^r}]$$

We additionally amended the metric of the CTT ability estimates (e.g., % correct) so this was more directly comparable to the IRT estimates. This was done by regression estimation with additional sets of simulated data - specifically we regressed *b*-parameters onto z_g indices, and used the average slope and intercept to rescale *p-values* on a psuedo-θ scale for graphing purposes.



Results

Visual ICC's generated from the CTT-parameters are being represented in Figure 2, with the IRT- 2PL derived ICC's of the same items being represented in Figure 3. The area between CTT-derived and IRT-derived ICC's (aka differential item functioning; "DIF") was calculated via the `geiger` package **(Harmon et al., 2020)**. The average plotting-space deviation for all 100 curves was 0.35 (*s* = .23). These differences were skewed (positively), reflecting that *most* of the DIF estimates were less than 0.35. Curves using both methodologies are very similar in shape and form, as we can see in the two items that we point out in each figure as well as three example items presented in the middle figure.

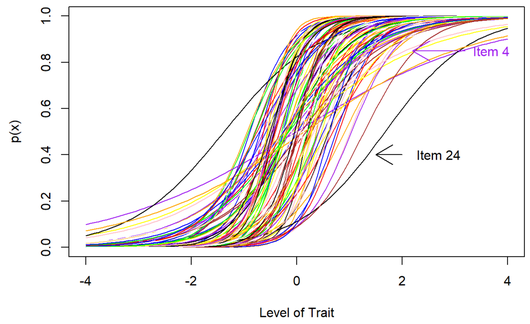


Figure 2: ICCs derived from only CTT parameters (with two noteworthy ICCs annotated).

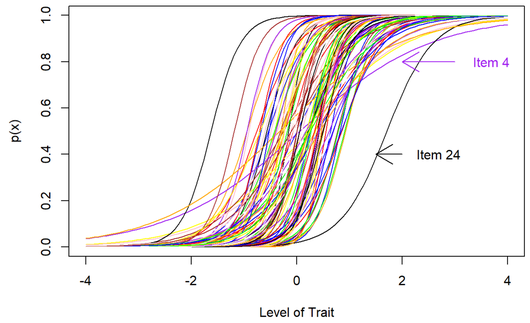


Figure 3: ICCs derived from IRT parameters (same noteworthy items annotated).

Discussion

Although invariance is a property historically associated exclusively for IRT applications, large scale data, truly random sampling, and large range items can also yield stable CTT item and person statistics **(Fan, 1998; Kulas et al., 2017)**. The current investigation is proof-of-concept that visual representations of CTT-derived item characteristics are feasible. We are currently working to extend our simulations and also finalize an `R` package that generates these CTT-derived ICCs. This poster was crafted via `posterdown` **(Thorne, 2019)**.

References

Chalmers, P. (2021). *Mirt: Multidimensional item response theory*. https://CRAN.R-project.org/package=mirt

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381.

Han, K. T. (2007). WinGen: Windows software that generates item response theory parameters and item responses. *Applied Psychological Measurement*, *31*(5), 457–459.

Harmon, L., Pennell, M., Brock, C., Brown, J., Challenger, W., Eastman, J., FitzJohn, R., Glor, R., Hunt, G., Revell, L., Slater, G., Uyeda, J., & Weir, J. (2020). *Geiger: Analysis of evolutionary diversification*. https://github.com/mwpennell/geiger-v2

Kulas, J. T., Smith, J. A., & Xu, H. (2017). Approximate functional relationship between irt and ctt item discrimination indices: A simulation, validation, and practical extension of lord's (1980) formula. *Journal of Applied Measurement*, *18*(4), 393–407.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Thorne, W. B. (2019). *Posterdown: An r package built to generate reproducible conference posters for the academic and professional world where powerpoint and pages just won't cut it.* https://github.com/brentthorne/posterdown