Item Characteristic Curve estimation from common Classical Test Theory indices

Diego Figueiras[1] & John T. Kulas[2]

[1] Montclair State University

[2] eRg

Author Note

Correspondence concerning this article should be addressed to Diego Figueiras, Dickson Hall 226. E-mail: figueirasd1@montclair.edu

Abstract

Item characteristic curves (ICC's) are graphical representations of important attributes of assessment items - most commonly *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. We propose an extension of this tradition of item characteristic visualization within the more commonly leveraged Classical Test Theory (CTT) framework. We first simulate binary (e.g., true *test*) data with varying item difficulty characteristics to generate empirically-derived linking coefficients between the IRT and CTT difficulty indices. The results of these simulations provided some degree of confidence regarding functional linking invariance. Next, we simulated datasets of varying item characteristic specification and generated ICCs derived from both IRT and CTT frameworks. Differential item functioning (DIF) was estimated by calculating the geometric area between the IRT- and CTT-derived ogives. The average DIF estimate was .2. Applying the CTT-derived ICCs to an applied sample of XXX test takers resulted in a mean DIF estimate of XXX. An `R` package, `ctticc`, performs the ICC calculations presented in the current paper and provides assessment specialists with visual representations of CTT-derived item characteristics.

Keywords: Classical Test Theory, Item Response Theory, item difficulty, item discrimination

Word count: X

Item Characteristic Curve estimation from common Classical Test Theory indices

Item characteristic curves are frequently consulted by psychometricians as visual indicators of important attributes of assessment items - most commonly *difficulty* and *discrimination*. Within these visual presentations the x-axis ranges along "trait" levels (by convention typically denoted with the greek θ), whereas the y-axis displays probabilities of responding to the item within a given response category. In the context of true tests, the response categories are binary[1], and the y-axis probability reflects the likelihood of a "correct" response[2]. Assessment specialists who consult ICC's usually do so from within the psychometric framework of either Item Response Theory (IRT) or Rasch modeling. These frameworks estimate the parameters necessary to plot the visual functions. Rasch models only estimate difficulty, and assume that differences in discrimination represent flaws in measurement. The IRT 2 parameter logistic model (2PL), however, estimates item discrimination in addition to item difficulty.

When interpreting an ICC, the observer extracts the relationship between a respondent's trait level and the expectation of answering the item correctly. If the curve transitions from low to high likelihood 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. Stated in the parlance of IRT or Rasch traditions, it does not take much θ to have a high likelihood of answering correctly. On the contrary, if the growth in the curve occurs primarily at higher trait levels, this indicates that the item is relatively more difficult. Through the lens of IRT, if discrimination is modeled and 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).

---

[1] With exception (see, for example, Masters, 1982; Muraki, 1997).

[2] Because the historical convention in test response is to code a correct response as "1" and an incorrect response as "0", the y-axis is commonly denoted as "$p(1)$" or "$p(1.0)$".

Item difficulty (the *b*-parameter) is scaled to the trait level associated with a 50% likelihood of correct response (e.g., it is scaled to θ). Item discrimination (the *a*-parameter) is the degree to which an item differentiates across individuals who are characterized as being relatively lower or higher on the trait and is scaled to the slope of the ICC function at the same 50% likelihood of correct response location[3]. 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.

Assessment specialists who calculate these CTT item indices don't typically (to our limited knowledge!) attempt to represent them visually, as is common in IRT and Rasch applications. However, ICC's based on CTT indices could possibly provide snapshot psychometric information as valuable as those gained from IRT- or Rasch-derived item parameters. The largest obstacle to psychometricians deeming CTT-derived visuals to be of value is likely tied to the concept of invariance, which refers to IRT parameter independence across item and person estimates. However, this property is often overstated, as invariance is only attained with perfect model-data fit (which never occurs), and is also only true after being subjected to linear transformation - commonly across samples (Rupp & Zumbo, 2006). Additionally, several comparative investigations have noted commonality between IRT and CTT difficulty and discrimination estimates as well as relative stability of CTT estimates when samples are large and/or judisciously constructed (Fan, 1998). Fan in fact summarizes that the IRT and CTT frameworks "...produce very similar item and person statistics" (p.379). Hambleton and Jones (1993) state that "no study provides enough empirical evidence on the extent of disparity between the two frameworks and the

---

[3] Within the 2PL. If more item characteristics are modeled, the *a*-parameter may be estimated at a different function location. ← Diego check this (look into *a*-paramter scaling for the 3PL; should be halfway between lower and upper asymptotes but I'm not 100% sure).

superiority of IRT over CTT despite the theoretical differences".

CTT and IRT Comparability Investigations

Fan (1998) examined correlations between CTT item statistics and the parameters derived from the three most popular IRT models (the 1-, 2-, and 3-parameter logistic). These correlations were very high, generally between .80 and .90. As for item discrimination, correlations were moderate to high, with only a few being very low[4]. Fan (1998) also investigated index invariance for all models. In theory, the major advantage of IRT models over CTT is that the latter has an interdependency between the item and person statistics, whereas under ideal circumstances IRT parameters have no such dependency. For example, within CTT examinations, the average item difficulty is equivalent to the average person score - these indices are merely reflective of averages computed across rows or columns. What Fan (1998) reported in his study, however, did not support the purported invariant advantage of IRT parameters over CTT indices. Both CTT-derived item difficulty and discrimination indices exhibited similar levels of invariance to the IRT-derived parameters, indicating that they were highly comparable.

Functional Relationship(s) between IRT and CTT Indices

Lord (1980) described a function to approximates the nonlinear relationship between the IRT a-parameter and the CTT discrimination index[5]:

---

[4] And in fact, as is presented below, the relationship between the IRT and CTT discrimination indices is non-linear - the Pearson's product moment correlation is therefore *not* the most appropriate index to capture the magnitude of this relationship.

[5] Lord (1980)'s CTT discrimination index is actually the item-test biserial correlation as opposed to the contemporarily more popular *corrected* item-total *point-biserial* correlation.

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

This formula wasn't intended for practical purposes but rather was specified in an attempt to help assessment specialists who were more familiar with CTT procedures to better understand the relationship to the IRT discrimination parameter. In an effort to move from the conceptual to a practical application, Kulas et al. (2017) proposed a modification that minimized the average residual (either a_i or r_i, where r_i is the *corrected* item-total *point-biserial* correlation).

The Kulas et al. (2017) investigations identified systematicly predictive differences in the relationship between a_i and r_i across items with differing item difficulty values, so their alteration to Lord (1980)'s formula included a qualifier effect for item difficulty (this formulaic specification is also retained in the current presentation):

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

Where g is the absolute deviation from 50% responding an item correctly and 50% responding incorrectly (e.g., a "p-value" of .5). z_g is the standard normal deviate associated with g. This transformation of the common p-value was recommended by Kulas et al. (2017) in order to scale the CTT index along a (closer to) interval-level metric more directly analogous to the IRT b-parameter. Figure 2 visualizes the re-specifications of Lord's formula at p-values (difficulty) of .5, .3 (or .7), and .1 (or .9) and highlights the nonlinear nature of this relationship - especially noticeable at high(er) levels of discrimination.

Study 1

Kulas et al. (2017) provided an extension of Lord (1980)'s conceptual formula, facilitating the scaling of the CTT corrected item-total correlation to the metric of the IRT

a-parameter. Fan (1998) demonstrated strong associations between the CTT p-value and IRT *b*-parameter, but did not attempt a scaling linkage. The ultimate goal of the current study is to generate CTT-derived ICCs. As a comparative standard, we also endeavor to compare the CTT-derived ICCs against IRT-derived ICCs. This comparison is only possible if the CTT statistic can be expressed on the IRT parameter metric (or vice versa). Study 1 therefore endeavors to develop a linking equation such that the CTT p-value may be translated to an IRT *b*-parameter metric. Specifically, we establish a regression equation predicting a "psuedo" b parameter from the z_g estimate (z_g is a derivative of the CTT p-value). The reason we're doing this is to get a better x-axis value for the ICCs (better means closer to the IRT-derived *b*).

Although the ogives could be specified directly from the CTT-derived statistics, we made a procedural decision to retain the IRT 2PL as our function specification:

$$P(\Theta) = \frac{1}{1 + e^{-1.7a(\Theta - b)}} \tag{3}$$

Our procedure therefore required the estimation of "pseudo" IRT parameters from the CTT indices. The *a* parameter was estimated via the formula specified in Kulas et al. (2017), while the *b* parameter was estimated via linking parameters identified via simulation.

1. Purpose: Getting a p-value → b-parameter linking equation [X]
2. Five different distributions of p-values (simulated data) [ ]
3. 10,000 runs each simulation (100 items, 10,000 "people" each) [ ]
4. Scrub extreme values (p-values essentially 0 and 1 need to be deleted) [ ]
5. Moderated regression to look for differences across p-value distribution [ ]
6. Tada! [ ]

Method

In order to generate a linking equation between CTT and IRT indices, we simulated datasets primarily differing in item difficulty. We kept the item set equal (100 items per simulation) and specified 5 different distributions of item difficulties. The first distribution was uniform, with p-values ranging from low to high at roughly equal levels of frequency. The second distribution was effectively normal with p-values, centered around 0.5. The third distribution was an inverted normal distribution also centered around 0.5. The fourth distribution was a negatively skewed distribution of p-values, and the fifth was positively skewed. Figure 3 provides a visual representation of the distributional forms that were desired in our simulations.

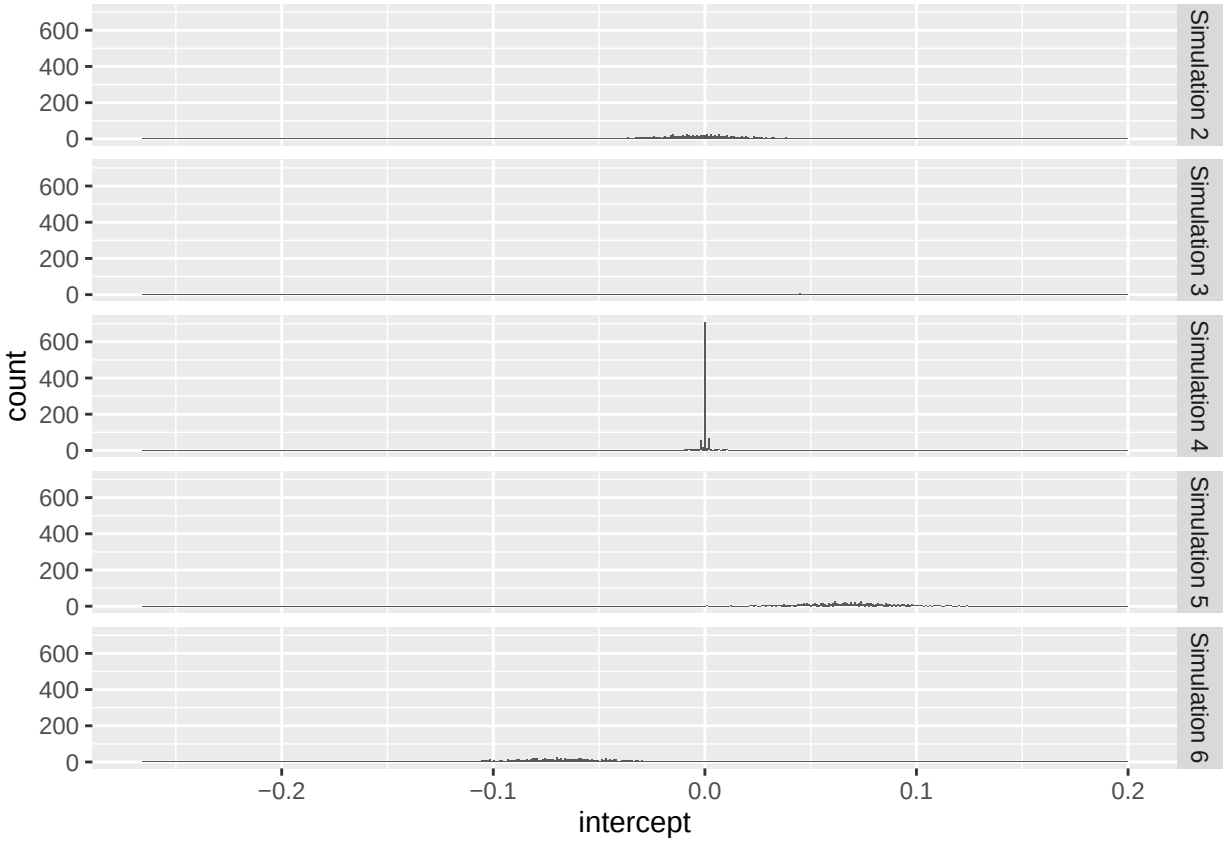
Then we computed regressions predicting the b-parameters using the standard normal deviate associated with the p-values on each simulation. The resulting regression coefficients for all simulations was approximately 2 and 0, indicating that our scaling was not sample dependent.

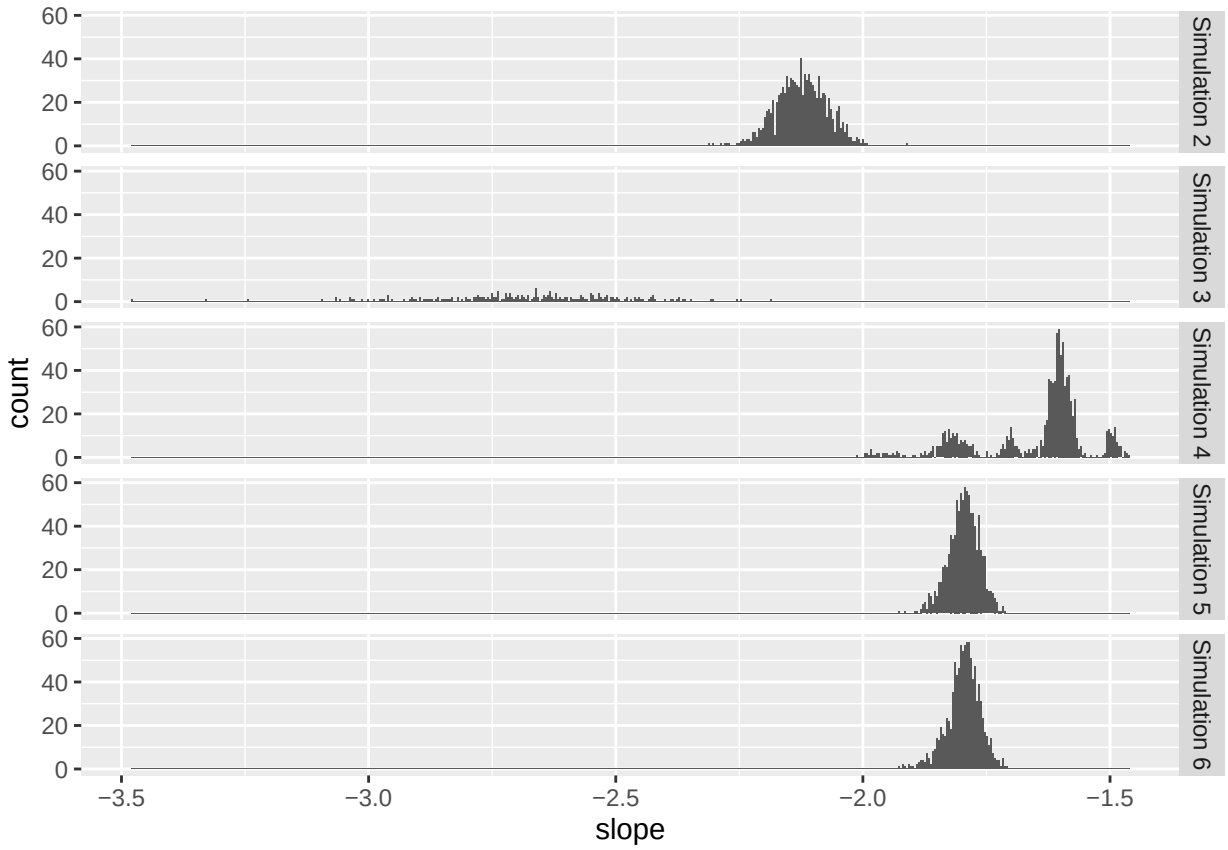
Procedure and methods

> *Note.* Maybe do a different linking via machine learning. Try to find the linking parameters (including p-value distributional shape and location) that minimize DIF across CTT and IRT ICCs (5/27/22 after unsuccessful Friday brainstorming especially regarding simulation 3 [the normally distributed p-values])

We simulated data using the WinGen program (Han, 2007). Our sample was 10,000 observations, with a mean of 0 and a standard deviation of 1. The number of items were 100, with response categories of either correct or incorrect (1 and 0). The mean for the a-parameter for the simulated data was 2, and the standard deviation 0.8. The mean for

the b-parameter was 0 and the standard deviation 0.5. The mirt package from (**R-mirt?**) was used to compute the IRT a-parameters and to plot the 2PL resulting model. As for the CTT-derived a-parameter, the modification to Lord (2012)'s formula described earlier was used, as well as the re-scaling for the p-values. We additionally changed the scale of the difficulty estimates of CTT so they were on the same scale as the IRT estimates. This was done by building a regression model using the CTT a-estimate to predict the IRT a-parameter. The resulting values from this model were used in plotting the CTT-derived ICC's. We tested the accuracy of this scale on six simulations of data, each with different shape and p-values. Each simulation consisted of 1,000 observations and 100 items. Simulation 1 was unidimensional, centered around a p-value of 0.5. Simulation 2... The results show that





Results

As shown by Figure 2, our plot looks very similar to that of (Kulas et al., 2017, p. 8). This confirms that our formula for computing the estimated a-parameter follows the exponential relationship we can see in (Kulas et al., 2017; Lord, 2012). Four random items were selected and plotted in Figure 3 using IRT and CTT-derived statistics. The blue curves were plotted using a IRT 2PL model, while the red curves were plotted with CTT-derived parameters.

Study 2 - Evaluating the Comparability of IRT and CTT ICC's

The purpose of study 2 is to simulate test data and generate ICC's based on the IRT model. Then we compare that to our CTT estimates and look at the differences. We

hypothesize that on average there won't be a big difference between the curves plotted with either methodology.

1. Use regression equation from Study 1
2. Many simulations
3. Compute DIF and report results
4. IF POSSIBLE, get real-world data and apply DIF algorithm (for example, reach out to ETS and ask for testing data)
5. Publish package (at least get to GitHub)
6. Tada!

Procedure and materials

The same simulated data as in study 1 was used. The mirt package from (**R-mirt?**) was used to compute and plot the IRT statistics. As we can see on Figure 3, the blue curves were plotted using 2PL IRT parameters (a and b), while the red curves were plotted using CTT parameters (p-values and corrected item-total correlations, re-scaling and modifying them with Kulas et al. (2017) formulas). To quantify the degree of difference between the two curves, the Area Between Curves was computed using (**R-geiger_a?**)'s package. This procedure was done for all 100 items.

Results

We used R (Version 4.2.1; R Core Team, 2022) and the R-packages *papaja* (Version 0.1.1; Aust & Barth, 2022), *psych* (Version 2.2.5; Revelle, 2022), *reticulate* (Version 1.25; Ushey et al., 2022), and *tinylabels* (Version 0.2.3; Barth, 2022) for all our analyses.

The area between ICC's was calculated between CTT-derived and IRT-derived ICC's. The average difference for all 100 curves was 0.35[6]. As we can see in Figure 4, most of the

[6] *Note.* Did the integral of the difference between the CTT and IRT functions using the "integrate"

data is skewed towards the lower end, indicating that out of the 100 items, most of them have areas between the curves of less than 0.35. For Figure 5 we plotted all the 100 ICC's that use CTT parameters, and for Figure 6 we did the same but with IRT parameters instead. 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.

Discussion

Important psychometric information can be gathered from ICC's, which are visual indicators typically of difficulty and discrimination. Psychometricians and other assessment specialists usually examine ICC's under the lenses of IRT and Rasch models. From a 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 CTT indices, but the most commonly calculated and consulted index is the corrected item-total correlation. Assessment specialists who consult these CTT parameters don't typically attempt to represent them visually, as is common in IRT and Rasch applications. However, there is perhaps little reason for them not to do so, as ICC's based on CTT parameters could provide snapshot psychometric information as valuable as those gained from IRT- or Rasch-derived ICC's. Here we first propose an application of ICC's with CTT indices, then we simulated data and quantified similarities and discrepancies between the IRT- and CTT-generated ICC's. Our hypothesis was that the Area Between Curves of these different ICC's would be small. Area between curves for 100 items was 0.35 on average. This result indicates that curves plotted with either IRT or CTT parameters show little difference. The nature of both models is mostly overlapping

---

function in the "stats" package (base R). Did a test to confirm this accurately reflects the area between curves by creating two curves, one with high discrimination and another with low discrimination, and seeing what the area between curves was using first the geiger package and then base R. Also roughly estimated by hand this diff. Base R seems to be the more accurate method.

when it comes to plotting visual representations such as ICC's. Practitioners and researchers that don't use IRT or Rasch models and instead opt to follow a CTT philosophy would benefit from having ICC's that use CTT statistics.

Of course there is always an intractability between the CTT item-difficulty index and respondent sample ability. The findings of previous comparison studies, however, point to the CTT estimates exhibiting some degree of invariance across respondent samples.

If this general idea is well-received (SIOP members would seem to represent a great barometer!) we would like to stress the CTT ICC's via further and more extensive conditions. That is, are there patterns that help explain CTT ICCs that diverge from their IRT counterparts? Although our simulations did generate a range of item difficulties and discriminations, we have not yet fully explored systematic patterns of extremely difficult/easy items as well as very poorly discriminating items. If patterns emerge, we would like to model predicted discrepancies via incorporating error bars within our visualizations.

represent some promise regarding plotted ICC's using IRT and CTT parameters. Our hypothesis was that the Area Between Curves of these different ICCs would be small. Area between curves for 100 items was 0.35 on average. This result indicates that curves plotted with either IRT or CTT parameters show little difference. The nature of both models is overlapping when it comes to plotting visual representations such as ICC's. Practitioners and researchers that don't use IRT or Rasch models and instead opt to follow a CTT philosophy would benefit from having ICC's that use CTT statistics.

Additionally, if there is interest in this general idea we would likely publish our function as a small `R` package, perhaps to supplement the `psych` package's "alpha" function, which produces corrected item-total correlations as well as p-values within the same output table (e.g., the "input" data is already available in tabular format).

References

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://github.com/crsh/papaja

Barth, M. (2022). *tinylabels: Lightweight variable labels*. https://cran.r-project.org/package=tinylabels

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.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

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

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 IRT to practical problems*. Hillsdale: Lawrence Erlbaum Associates.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

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.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*,

47(2), 149–174.

Muraki, E. (1997). A generalized partial credit model. In *Handbook of modern item response theory* (pp. 153–164). Springer.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rupp, A. A., & Zumbo, B. D. (2004). A note on how to quantify and report whether IRT parameter invariance holds: When pearson correlations are not enough. *Educational and Psychological Measurement*, *64*(4), 588–599.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63–84.

Ushey, K., Allaire, J., & Tang, Y. (2022). *Reticulate: Interface to 'python'*. https://CRAN.R-project.org/package=reticulate

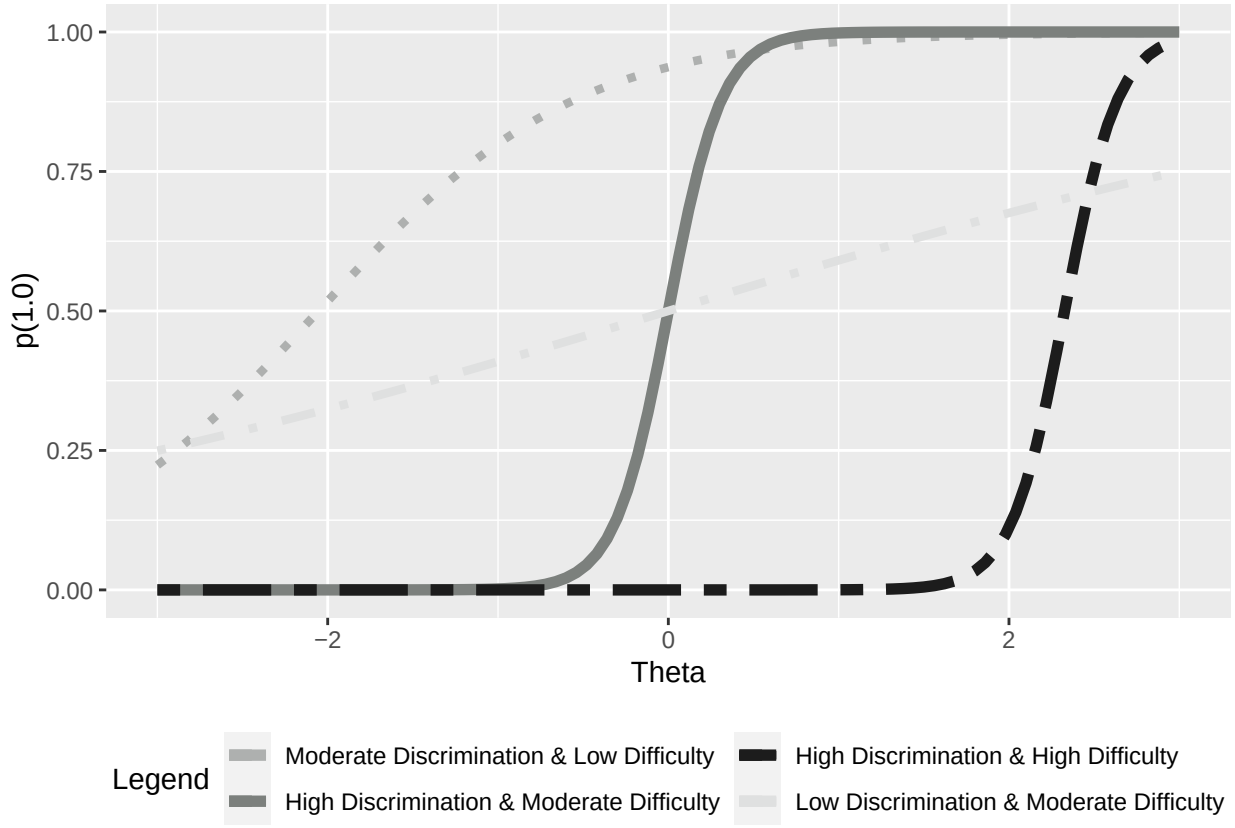


Figure 1. Item characteristic curves demonstrating differences in item difficulty and discrimination.

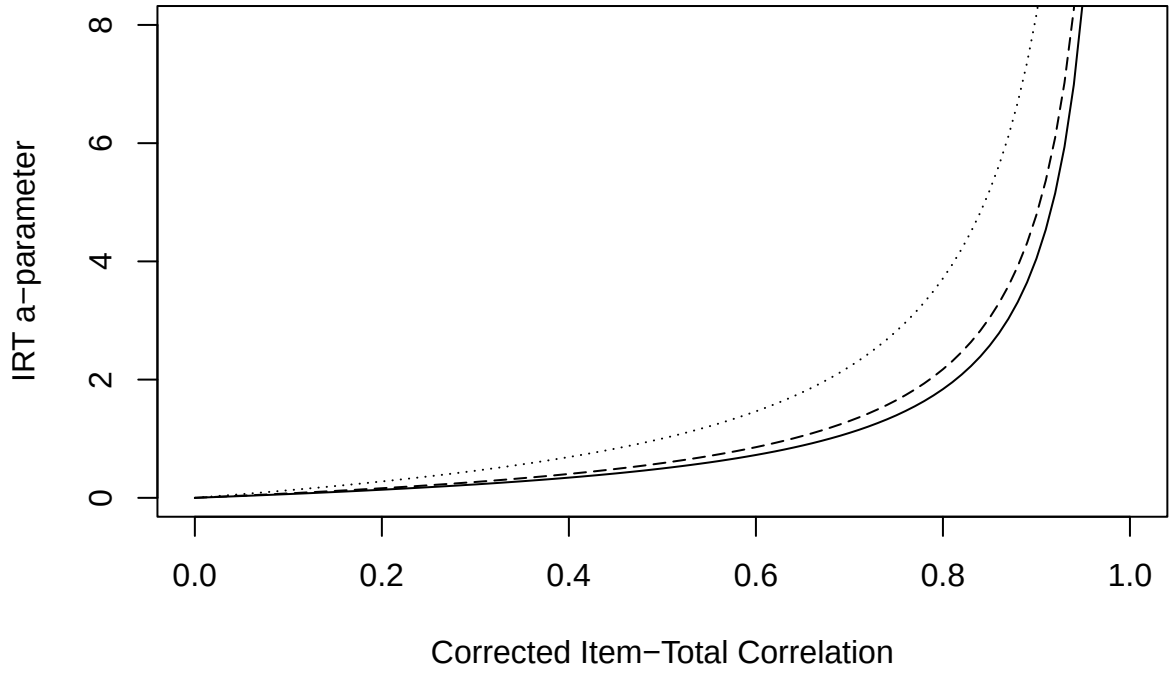


Figure 2. Kulas et al. (2017) functional relationship between the IRT *a* parameter and the CTT corrected-item total correlation as a function of item difficulty (p-value; solid = .5, dashed = .3/.7, dotted = .1/.9).

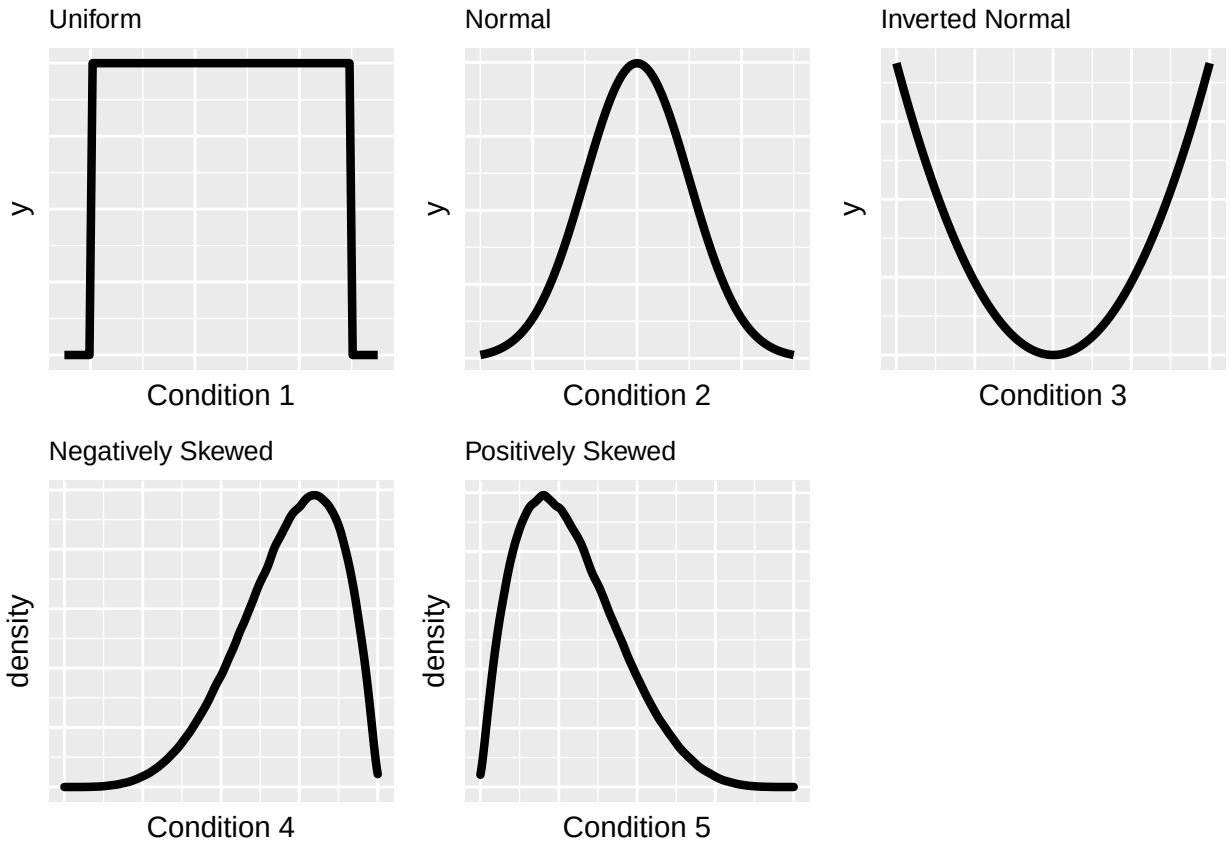


Figure 3. Shape of prescribed distributions of p-values across Study 1 conditions.

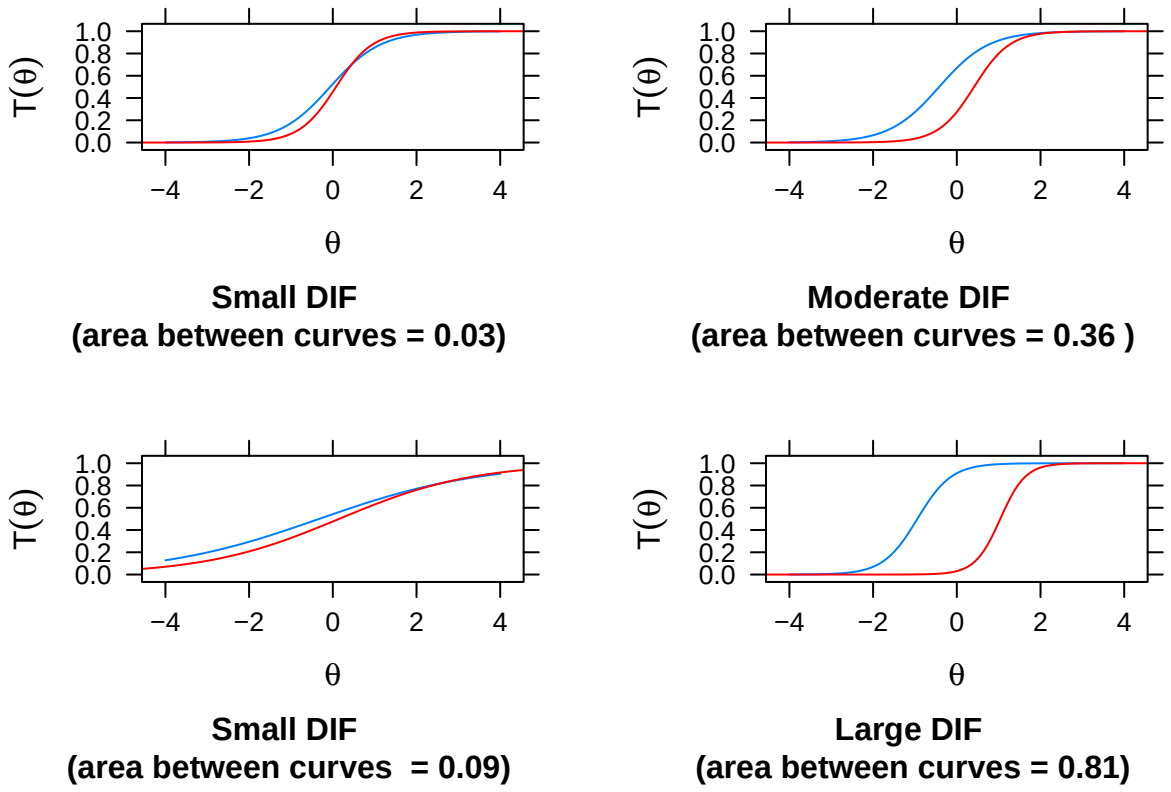


Figure 4. Four ICCs highlighting the difference between CTT and IRT-derivated ICCs at different levels of DIF.

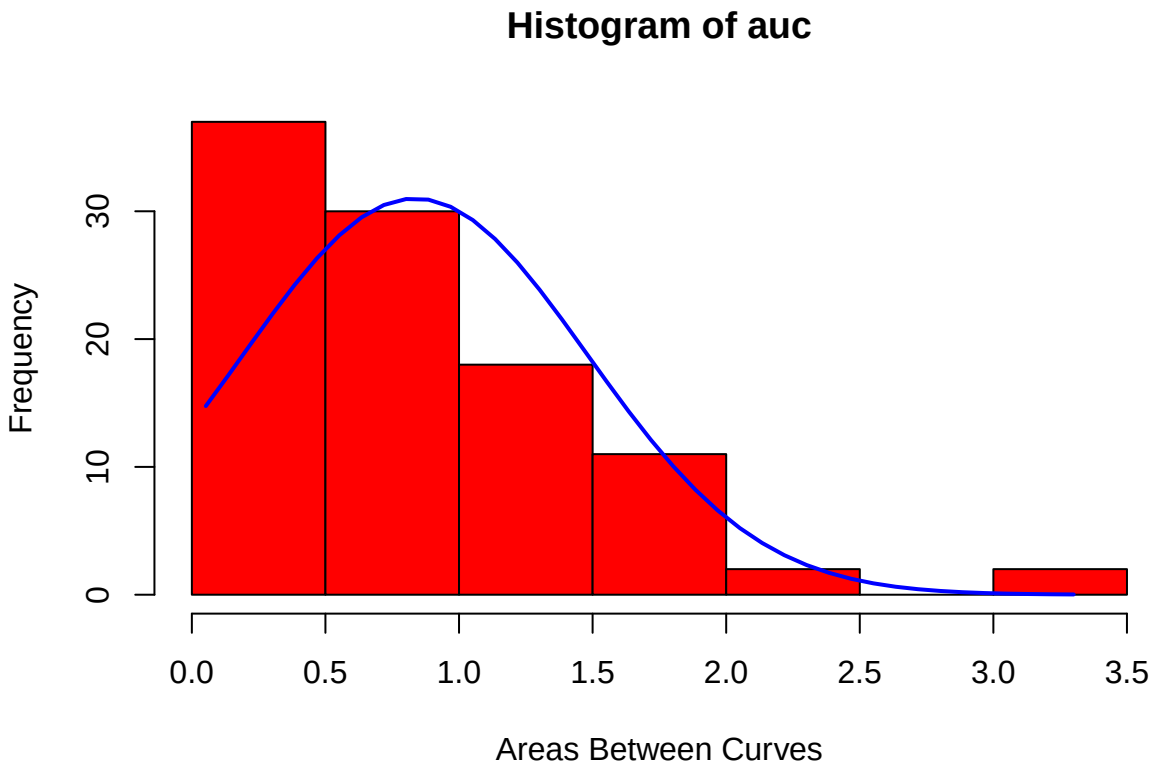


Figure 5. Histogram of all areas between ICCs plotted using IRT parameters vs ICCs plotted using CTT parameters.

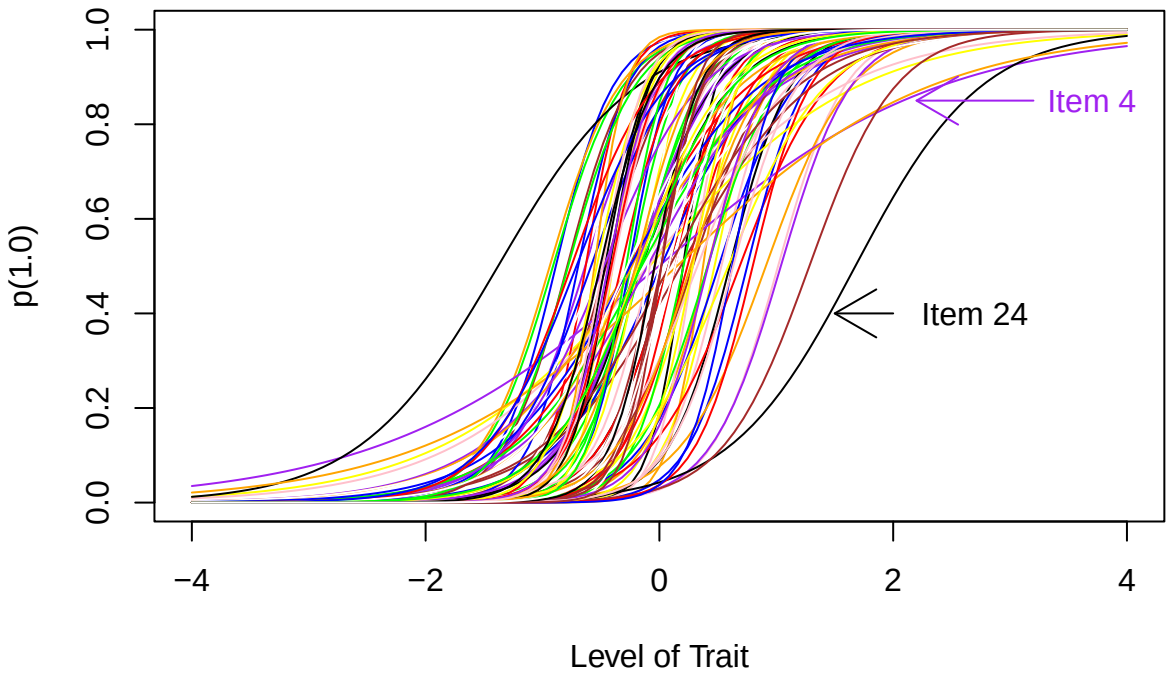


Figure 6. ICCs derived from only CTT parameters (with two noteworthy ICCs annotated).

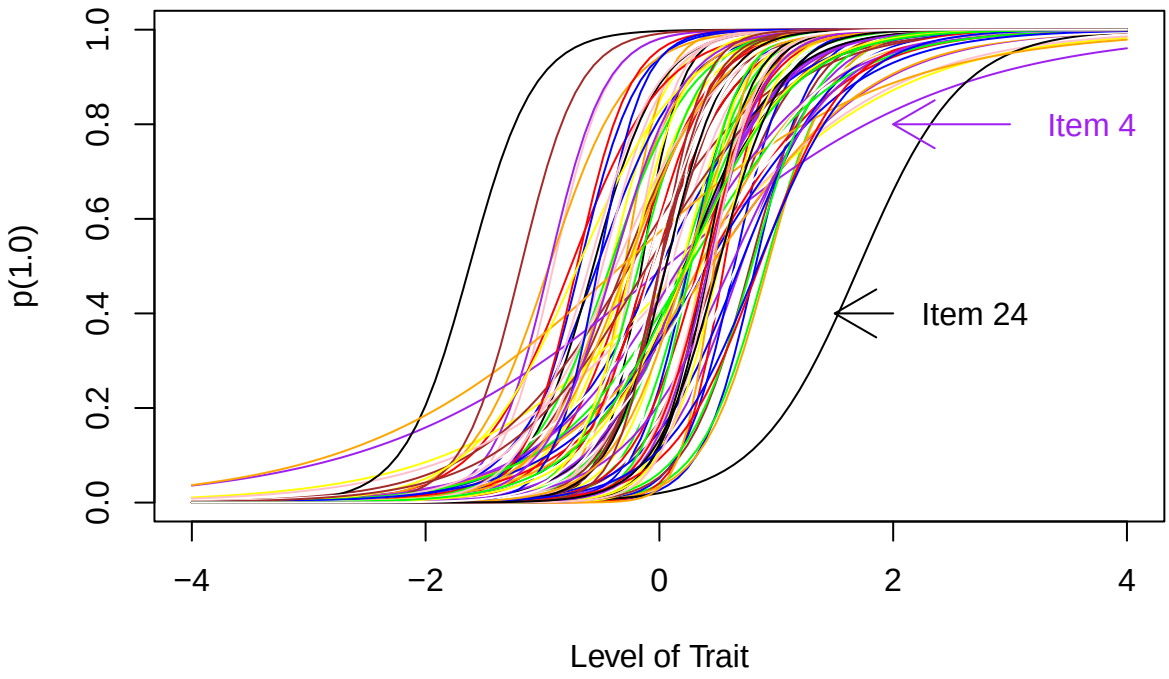


Figure 7. Typical ICCs derived from IRT parameters (same noteworthy items annotated).

Appendix

Cut stuff

There have also been suggestions that the invariance property be conceptualized as a graded continuum instead of a categorical (invariant or non-invariant) population property (Hambleton et al., 1991; Rupp & Zumbo, 2004). Estimates of IRT parameters across different calibration runs can be looked at for evidence of a possible lack of invariance. This doesn't happen with CTT item parameters, since they will always be sample-dependent. This dependency, however, is greatly influenced by the sampling strategy. Large scale data, truly random sampling, and large range items could give comparable CTT item and person statistics across testing populations and occasions (Kulas et al., 2017). Additionally, there are several empirical investigations that note high levels of "invariance" of CTT estimates, in some cases surpassing IRT item estimates in their capacity to have cross-sample stability (Fan, 1998; Macdonald & Paunonen, 2002).

An adjustment to Lord (2012)'s formula giving the functional relationship between the "non-invariant" CTT and "invariant" IRT statistics becomes useful in comparing the two methodologies, despite the supposed lack of invariance from CTT. So even though here we acknowledge that invariance is a categorical IRT property, we still follow the functional modification proposed by Kulas et al. (2017), noting that having a large sample that is truly random and whose items are normally distributed and have a center at the moderate difficulty can help reduce threats to CTT "invariance".