

Item Characteristic Curve estimation via Classical Test Theory specification

Diego Figueiras¹ & John T. Kulas²

¹ Montclair State University

² eRg

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 derive linking coefficients between the IRT and CTT difficulty and discrimination indices. The results of these simulations provided some degree of confidence regarding functional linking coefficient invariance. Next, we simulated a sample test dataset 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 low within this simulated dataset ($\overline{DIF} = .08$ on our 13x1 dimensional plotting space). Applying the CTT-derived ICCs to six different applied tests of 20,000 real-life examinees resulted in a comparable mean DIF estimate of .12. Collectively, these results should provide some confidence to test specialists interested in creating visual representations of CTT-derived item characteristics. An R package, `ctticc`, performs the ICC calculations presented in the current paper and generates ICC plots directly from CTT indices.

Keywords: Classical Test Theory, Item Response Theory, item difficulty, item discrimination
Word count: X

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 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 binary¹, and the y-axis probability reflects the likelihood of a “correct” response². Assessment specialists who consult ICC's usually do so from within the psychometric framework of either Item Response Theory (IRT) or Rasch modeling. These approaches estimate the item characteristics that define the visual functions.

Rasch models only estimate difficulty, and assume that differences in discrimination represent flaws in measurement (e.g., Wright, 1977). The IRT 2 parameter logistic (2PL) and higher order models, however, estimate item discrimination in addition to item difficulty.

When interpreting an ICC representing a true test item, the observer extracts the relationship between a respondent's trait level and the corresponding expectation of answering the item correctly. If the function 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

Materials for this research were provided by Educational Testing Service (ETS) and the TOEFL program. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researcher(s).

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

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

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

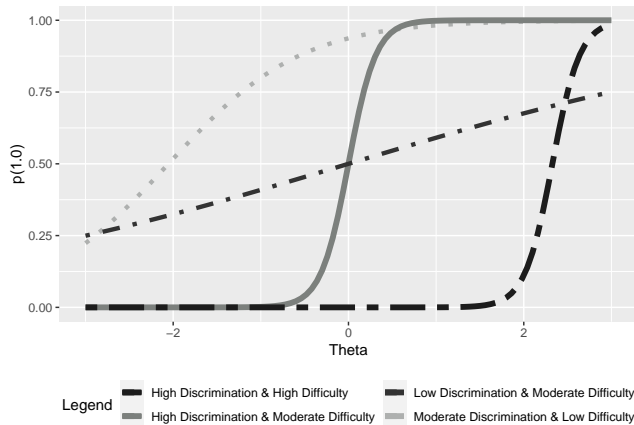


Figure 1. Item characteristic curves demonstrating differences in item difficulty and discrimination.

and the curve is sharp (e.g., strongly vertical), this indicates greater item discrimination across trait levels; if it is flatter, that is an indication of poorer discrimination (see Figure 1 for some exemplar ICCs).

Item difficulty (the IRT b -parameter) is typically expressed as the trait level associated with a 50% likelihood of correct response (e.g., it is scaled to θ). Item discrimination (the a -parameter) reflects the degree to which an item differentiates across individuals who are located 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 location³. 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”)⁴. Item discrimination can be conveyed via a few different CTT indices, but the most commonly calculated and consulted contemporary index is the corrected item-total correlation.

Assessment specialists who calculate these CTT item indices do not, by tradition, additionally represent them visually, as is common in IRT and Rasch applications. However, ICCs based on CTT indices should provide snapshot psychometric information comparably as valuable as those conveyed by 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 is never attained), and is also only true after being subjected to linear transformation - commonly across samples (Rupp & Zumbo, 2006).⁵ Additionally, several comparative investigations have noted commonality between IRT and CTT difficulty and discrimination estimates as well as absolute sta-

bility of CTT estimates (e.g., Fan, 1998; Kulas et al., 2017; Lawson, 1991).

CTT and IRT Comparability Investigations

Fan (1998) examined associations between CTT item statistics and the parameters derived from the three most popular IRT models (the 1-, 2-, and 3-parameter logistic). Correlations were very high for difficulty estimates - generally between .80 and .90. These findings converged with both earlier and later investigations that also found strong correspondence between difficulty estimates (e.g., Lawson, 1991; Macdonald & Paunonen, 2002). As for item discrimination, correlations were *moderate* to high, with only a few being very low - these estimates tended to be even poorer in Macdonald and Paunonen (2002)’s Monte Carlo investigation⁶.

Fan (1998) also investigated index invariance for all models. In theory, the primary advantage of IRT- over CTT-models is that the latter has an intractable dependency between the item and person statistics, whereas under ideal circumstances IRT parameters have no such dependency. Within CTT examinations, for example, the average item difficulty is necessarily equivalent to the average person score - these CTT 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. Fan (1998) in fact summarizes that the IRT and CTT frameworks “... produce very similar item and person statistics” (p.379). Macdonald and Paunonen (2002)’s Monte Carlo simulations agreed with the Fan (1998) conclusion regarding *difficulty* and person estimates, but did note superior performance of IRT relative to CTT discrimination estimates.

³Within the 2PL. If additional item characteristics are modeled, the a -parameter may be estimated at a different function location.

⁴Without being provided additional context, the psychometric “ p value” is only distinguishable from the inferential statistic “ p -value” via conventional notation adherence: omission of the grammatical dash in the case of the difficulty index.

⁵There 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). If this lens is adopted, the concept of “acceptable” levels of invariance further augments the rationale behind the construction of CTT-derived ICCs.

⁶As is presented below, the relationship between the IRT and CTT discrimination indices is non-linear. The Pearson’s product moment correlation was consulted in both Fan (1998) and Macdonald and Paunonen (2002) although it is not the most appropriate index to capture the magnitude of the relationship.

Relationship(s) between IRT and CTT Indices

In addition to the comparability studies, there have been some investigations attempting to model direct associations between IRT and CTT indices. Lord (1980) first provided a conceptual function to approximate the nonlinear relationship between the IRT a -parameter and the CTT discrimination index⁷:

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

This formula was not intended for practical applications but was rather presented as an attempt to help assessment specialists who were more familiar with CTT procedures to better understand the IRT discrimination parameter. In an effort to move from the conceptual to a more practical application, Kulas et al. (2017) proposed a modification focused on minimizing predicted residual values (the predicted a_i).

The Kulas et al. (2017) investigations identified systematically 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 moderating effect for item difficulty, with r_i also being operationalized as the *point-biserial* correlation between an item's binary response and the *corrected* total test score:

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

Within formula (2), g represents the absolute deviation from 50% responding to 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 presents a visual representation of the exponential relationship between the two models' discrimination indices. We retained the Kulas et al. (2017) \hat{a}_i and z_g indices as "starting points" in the current investigation.

Summary and Overall Purpose

The primary goal of the current project was to generate CTT-derived ICCs. As a standard of comparison, however, we also endeavored to evaluate the CTT-derived ICCs against their IRT-derived counterparts. These comparisons are only feasible if the CTT indices can be reasonably expressed on the IRT parameter metric (or vice versa). Fan (1998) demonstrated strong associations between the CTT p value and IRT

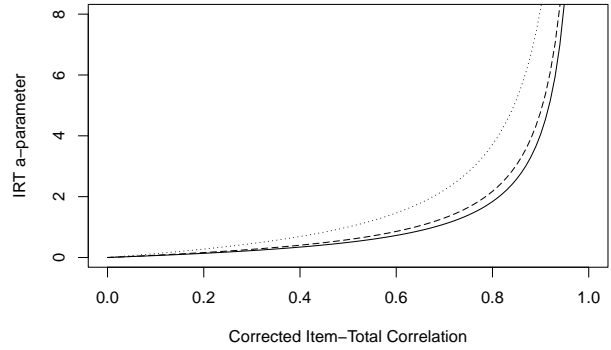


Figure 2. Kulas et al. (2017)'s proposed functional relationship between the IRT a parameter and the CTT corrected-item total correlation as a function of item difficulty (p value; solid = .5, dashed = .3/.7, dotted = .1/.9).

b -parameter, but did not attempt a scaling linkage. Similarly, Kulas et al. (2017) focused on nonlinear functional specification rather than metric of expression. Study 1 is therefore focused on the development of linking equations such that the CTT p value and corrected item-total correlation may be approximated along the IRT b - and a -parameter metrics.

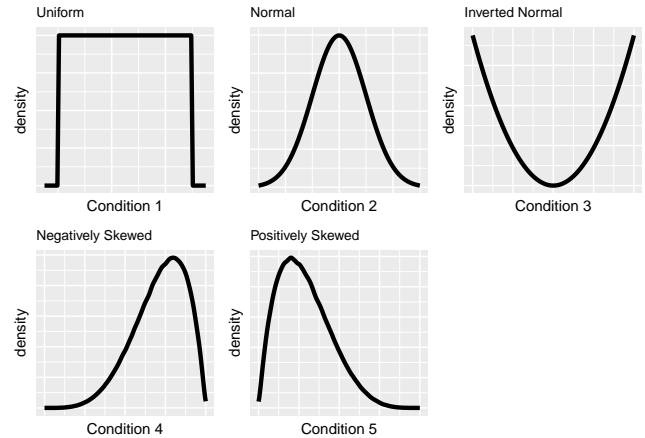


Figure 3. Shape of prescribed distributions of p values across Study 1 conditions.

Study 1

Procedure and methods

Study 1 focused on simulated datasets of binary item responses. The simulated data prescriptively differed in distri-

⁷Lord (1980)'s CTT discrimination index is the item-test biserial correlation as opposed to the contemporarily more popular *corrected* item-total *point-biserial* correlation.

butions of item difficulty while keeping the numbers of items ($k=100$) and “respondents” ($n=10,000$) equivalent. The first distributional form was uniform, with p values ranging from low (approaching 0) to high (approaching 1) at roughly equal levels of frequency. The second distribution was effectively normal with p values centered around 0.5. The third set of distributions was inverted normal and also centered around 0.5. The fourth distributional form was negatively skewed, and the fifth was positively skewed. Figure 3 provides a visual representation of idealized distributional forms that were prescribed across our simulations

For each simulation, we estimated CTT p values and corrected item-total correlations via the `psych` package (William Revelle, 2023). The 2PL was also applied via the `mirt` package (Chalmers, 2012), and a and b parameters were extracted. Regressions were applied within each simulation to predict the IRT b parameter from the p value derived z_g statistic. The simulated data was derived by first specifying distributions of p values (via the `runif()` base-R function for the uniform and inverted distribution conditions, and `rnorm()` for the normal and skewed distribution conditions). p values within each simulation were then referenced to generate binary item responses via application of the Chalmers (2012) `simIrt()` function.

Across all simulations, for items that realized extreme empirical p values (less than 0.02 or greater than 0.98), 200 responses were modified. For items with p values less than 0.02, 200 random responses of “1” were substituted. For items with p values greater than 0.98, 200 random responses values of “0” were imputed. This was done so the IRT models would be less likely converge on disproportionately extreme parameter estimates.

Across both Study 1 and 2, all analyses and manuscript development was accomplished via R (Version 4.2.3; R Core Team, 2023) and the R-packages *ape* (Version 5.7.1; Paradis & Schliep, 2019), *cticc* (Version 0.1.0; Figueiras & Kulas, 2023), *descr* (Version 1.1.7; Dirk Enzmann et al., 2023), *dplyr* (Version 1.1.1; Wickham, François, et al., 2023), *forcats* (Version 1.0.0; Wickham, 2023), *geiger* (Version 2.0.11; Alfaro et al., 2009; Eastman et al., 2011; Harmon et al., 2008; Pennell et al., 2014; Slater et al., 2012), *ggplot2* (Version 3.4.2; Wickham, 2016), *ggthemes* (Version 4.2.4; Arnold, 2021), *gridExtra* (Version 2.3; Auguie, 2017), *lattice* (Version 0.21.8; Sarkar, 2008; Sarkar & Andrews, 2022), *latticeExtra* (Version 0.6.30; Sarkar & Andrews, 2022), *lubridate* (Version 1.9.2; Golemund & Wickham, 2011), *maps* (Version 3.4.1; Richard A. Becker et al., 2022), *mirt* (Version 1.39; Chalmers, 2012), *papaja* (Version 0.1.1; Aust & Barth, 2022), *phytools* (Version 1.9.16; Revell, 2012), *plotly* (Version 4.10.2; Sievert, 2020), *psych* (Version 2.3.6; William Revelle, 2023), *purrr* (Version 1.0.1; Wickham & Henry, 2023), *readr* (Version 2.1.4; Wickham, Hester, et al., 2023),

readxl (Version 1.4.2; Wickham & Bryan, 2023), *reticulate* (Version 1.30; Ushey et al., 2023), *scales* (Version 1.2.1; Wickham & Seidel, 2022), *stringr* (Version 1.5.0; Wickham, 2022), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidyr* (Version 1.3.0; Wickham, Vaughan, et al., 2023), *tidyverse* (Version 2.0.0; Wickham et al., 2019), *tinylab* (Version 0.2.3; Barth, 2022), *viridis* (Version 0.6.4; Garnier et al., 2023a, 2023b), and *viridisLite* (Version 0.4.2; Garnier et al., 2023b).

Results

For all reported analyses, items that evidenced b values more extreme than $|3|$ were excluded. We made this procedural decision because our primary interest was in placing the CTT estimates on scales most likely approximating the IRT metric. Extreme cases would have the potential to skew linking coefficients.

Figure 5 shows the distribution of five million regression slopes and intercepts estimated across our simulations. A regression predicting b from z_g returned a R^2 value of 0.96. Regarding possible differences across simulation conditions, a moderated hierarchical regression did yield a significant interaction effect ($F_{(1,4,943,349)} = 1,783.21, p < .001$), but this was due to our large sample size ($\Delta R^2 = 0.00$).

We first noted across conditions that the Kulas et al. (2017) \hat{a}_i was systematically underpredicting the IRT a -parameter. Regressions to further modify the \hat{a}_i scaling resulted in the specification of a slope coefficient modifying value of 1.72, which was retained in all subsequent analyses. We retained the slope and intercept of this regression equation to inform our z_g scaling coefficients, which were $b = 1.54$ and $a = 0.00$. Application of these values to the z_g specification theoretically places the CTT estimate on a metric more closely approximating the IRT parameter. Study 2 tests this premise empirically.

Across all five conditions, simulated distributions resulted in an average empirical a -estimate of 1.50 ($sd = 0.50$) and average empirical b -estimate of 0.00 ($sd = 1.17$). The average z_g was 0.00 ($sd = 0.75$)⁸, and the average \hat{a}_i was 1.46 ($sd = 0.43$). A paired samples t-test revealed a consistent underprediction effect for \hat{a}_i relative to the IRT a -parameter ($\bar{D} = 0.04; t_{(4,940,181)} = 613.01, p < 0.001$). Visual inspection of this effect confirmed that the underestimation became more likely with strongly discriminating items, and the effect was universal (e.g., systematic - there is likely further refinement that could be applied to the \hat{a}_i formula in future applications). The average difference between the IRT b -parameter and z_g was non-significant ($\bar{D} = 0.00; t_{(4,940,181)} = -1.34, p = 0.18$).

⁸ z_g specification was simplified to the inverse of the standard normal deviate in all current paper analyses.

Study 2

Procedure and Methods

The purpose of Study 2 was to evaluate the comparability of IRT- and CTT-derived ICCs. We generated ICCs from one simulated and six real-world test datasets. The real-world datasets represent responses from the Test of English as a Foreign Language institutional testing program (TOEFL ITP). The TOEFL ITP has subscales of: reading ($k=39$), listening ($k=40$), and speaking ($k=35$). There were two different test forms. Datasets representing responses to items from the two different forms defining the three subscales both included responses from 10,000 examinees, and the examinee samples for the two forms did not overlap.

The simulated data were generated using Wingen (Han, 2007). One dataset with 100 binary response items and 10,000 “respondents” was requested. We generated responses derived from an effectively normal and centrally located fictional ability distribution and items with a mean a -parameter value of 2 (sd = 0.8) and a mean b -parameter value of 0 (sd = 0.5).

Differential item functioning (DIF) was estimated via directly calculating the area between ICCs. This number reflects a two-dimensional plotting space defined by an x-axis ranging from -6 to 6 and a y-axis ranging from 0 to 1, resulting a maximal possible 2-dimensional geometric area of 13. This estimation is sensitive to both uniform as well as non-uniform DIF.

Results

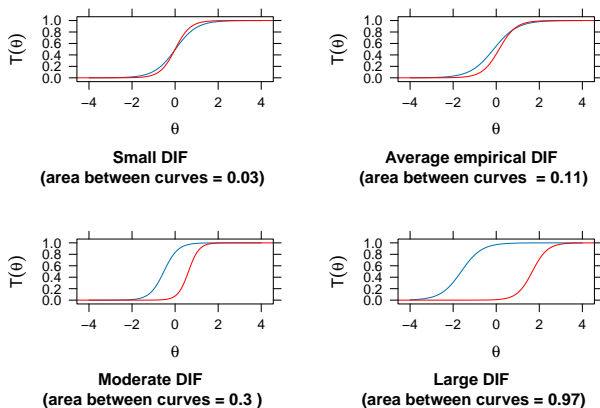


Figure 4. Four ICCs highlighting the difference between CTT and IRT-derived ICCs at different levels of DIF.

The mirt package (Chalmers, 2012) was again retained for 2PL estimation. Across all 7 datasets, DIF ranged from a smallest value of 0.00 to a greatest value of 1.17. Figure 6

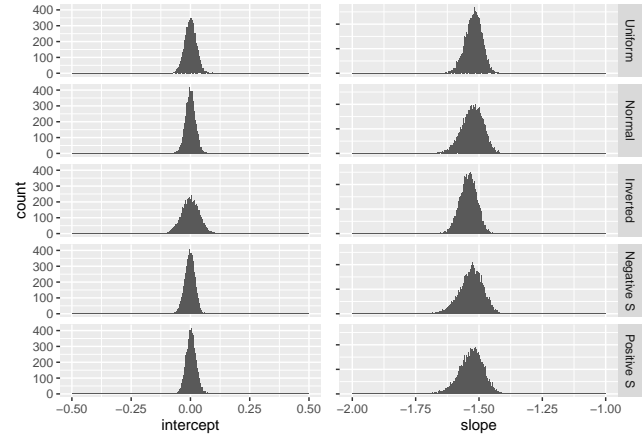


Figure 5. Individual intercepts and slopes grouped by study 2 simulation.

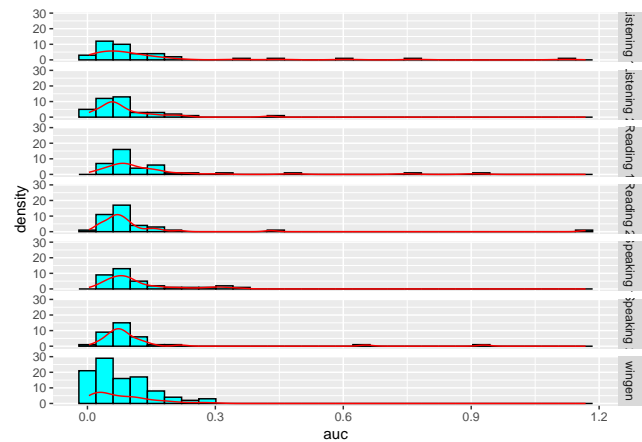


Figure 6. Histogram of geometric areas between ICCs plotted with IRT parameters versus those plotted with CTT statistics.

presents histograms of the individual DIF magnitudes organized by focal test. The average DIF across all 7 datasets was 0.11 (sd=0.14). There was no difference in average DIF across the 7 datasets ($F_{(6,321)}=2.04, p > .05$). Only 17 (5.2%) of the 328 total investigated items exhibited a DIF above 0.3. The simulated test data returned an average DIF estimate of 0.08 whereas the TOEFL tests returned an average estimate of 0.12 across 228 investigated items.

Figure 4 presents ICCs representing items that exhibited small, empirically average, moderate, and notably large degrees of DIF (“small”, “moderate”, and “large” are subjective author specifications based on the omnibus distribution of DIF across all 7 tests). Here, 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-scaled with Study 1 modified Kulas et al., 2017’s formulas).

Discussion

Valuable psychometric information can be quickly extracted from ICC's. Assessment specialists, by tradition, exclusively seek to generate ICC's within the frameworks of IRT and/or Rasch models. Of course there will always be an intractability between the CTT item indices and respondent sample abilities. The findings of previous comparison studies, however, point to the CTT estimates exhibiting a moderate-to-high degree of: 1) comparability to IRT parameters, and 2) invariance across respondent samples. The proposal of the current presentation is that ICC's derived from CTT statistics may provide snapshot psychometric information similar in value to those derived from IRT parameters. There's little *practical* (as well as perhaps empirical) reason for assessment specialists who rely on a CTT framework to be denied the privilege of ICCs.

The noted geometric areas between curves for all investigated items was, on average, low. This represents 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. The current specifications are available to apply via a small R package called `ctticc`. Because the ICCs are derived from the same individuals, the current application does not require the same parameter scaling that is typically required in DIF investigations where the ICCs reflect responses from different groups that likely have different underlying distributions of ability.

CTT item statistics 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).

Future improvements could stress the CTT ICC's via further and more extensive simulations. That is, are there patterns that help isolate the CTT ICCs that diverge from the IRT-derived ICCs? 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.

Visual inspection of the current study indices were consistent in that items with a -parameters above 1 occasionally had consistently underestimated \hat{a} values. Z_g is also consistently overpredicting at extreme values (regardless of whether it was an extreme positive or an extreme negative, the pattern is consistent). Further refinement can be made. If patterns emerge, we would like to model predicted discrepancies via incorporating error bars within our visualizations. Although scaled inventory responses are more common in Psychological assessment applications, We do not believe a visual representation of the polytomous item response function (IRF) would be as practically informative, and do not foresee extensions to inventory response.

Common inferential DIF indices such as likelihood ratio test Thissen et al. (n.d.) or Lord's chi-square test Lord (1980) require the specification of parameter variance and covariance estimates, which were not sought in the case of the CTT indices. Future investigations may wish to attempt an application of inferential DIF indices to obtain estimates more comparable to others found in the published and commercial literatures.

IRT analyses also require large samples of data. These CTT-derived ICC estimates may be useful to individuals who wish to ultimately apply IRT, but are limited in their data sample size, as is the case for example in classroom settings.

IRT models can converge with wildly large b -parameter estimates with extremely difficult or easy items Baker and Kim (2004). This is possibly the reason previous investigations have noted an advantage to the CTT-derived difficulty estimates with regard to index invariance (e.g., Fan, 1998; Lawson, 1991). The IRT estimate is bound by positive and negative infinity - the CTT estimate is finite and does not suffer this possibility.

There is a popular Thorndike (1982) quote regarding the (at the time) rising popularity of IRT models that has been noted in previous IRT-CTT investigations (e.g., Fan, 1998; Macdonald & Paunonen, 2002). The quote also bears inclusion here:

For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. (p. 12)

We agree, and offer a gift to CTT practitioners from the IRT tradition - not

References

- Alfaro, M. E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D. L., Carnevale, G., & Harmon, L. J. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *PNAS*, *106*, 13410–13414. <https://doi.org/10.1073/pnas.0811087106>
- Arnold, J. B. (2021). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*. <https://CRAN.R-project.org/package=ggthemes>
- Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. <https://CRAN.R-project.org/package=gridExtra>
- Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown*. <https://github.com/crsh/papaja>
- Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. CRC press.
- Barth, M. (2022). *tinylabls: Lightweight variable labels*. <https://cran.r-project.org/package=tinylabls>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Dirk Enzmann, J. Aquino. I. R. source code and/or documentation written by, Schwartz, M., Jain, N., & Kraft, S. (2023). *Descr: Descriptive statistics*. <https://CRAN.R-project.org/package=descr>
- Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., & Harmon, L. J. (2011). A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, *65*, 3578–3589. <https://doi.org/10.1111/j.1558-5646.2011.01401.x>
- 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.
- Figueiras, D., & Kulas, J. (2023). *Ctticc: Item characteristic curve estimation from classical test theory indices*.
- Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro, A., Sciaini, Marco, Scherer, & Cédric. (2023a). *viridis(Lite) - colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4679423>
- Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro, A., Sciaini, Marco, Scherer, & Cédric. (2023b). *viridis(Lite) - colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4678327>
- Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. <https://www.jstatsoft.org/v40/i03/>
- 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.
- Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*, 129–131. <https://doi.org/10.1093/bioinformatics/btm538>
- 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.
- Lawson, S. (1991). *One parameter latent trait measurement: Do the results justify the effort?* (B. Thompson, Ed.; Vol. 1, pp. 159–168). JAI.
- Lord, F. M. (1980). *Applications of IRT to practical problems*. Hillsdale: Lawrence Erlbaum Associates.
- 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.
- Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. <https://CRAN.R-project.org/package=tibble>
- Muraki, E. (1997). A generalized partial credit model. In *Handbook of modern item response theory* (pp. 153–164). Springer.
- Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., Fitzjohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic

- trees. *Bioinformatics*, 30, 2216–2218. <https://doi.org/10.1093/bioinformatics/btu181>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revell, L. J. (2012). Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3, 217–223. <https://doi.org/10.1111/j.2041-210X.2011.00169.x>
- Richard A. Becker, O. S. code by, Ray Brownrigg. Enhancements by Thomas P Minka, A. R. Wilks. R. version by, & Deckmyn., A. (2022). *Maps: Draw geographical maps*. <https://CRAN.R-project.org/package=maps>
- 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.
- Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. Springer. <http://lmdvr.r-forge.r-project.org>
- Sarkar, D., & Andrews, F. (2022). *latticeExtra: Extra graphical utilities based on lattice*. <https://CRAN.R-project.org/package=latticeExtra>
- Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny*. Chapman; Hall/CRC. <https://plotly-r.com>
- Slater, G. J., Harmon, L. J., Wegmann, D., Joyce, P., Revell, L. J., & Alfaro, M. E. (2012). Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution*, 66, 752–762. <https://doi.org/10.1111/j.1558-5646.2011.01474.x>
- Thissen, D., Steinberg, L., & Wainer, H. (n.d.). *Test validity* (H. Wainer & H. Braun, Eds.; pp. 147–170). Lawrence Erlbaum.
- Thorndike, R. L. (1982). *Educational measurement: Theory and practice* (D. Spearritt, Ed.; pp. 3–13). ERIC Clearinghouse of Tests, Measurements; Evaluations.
- Ushey, K., Allaire, J., & Tang, Y. (2023). *Reticulate: Interface to 'python'*. <https://CRAN.R-project.org/package=reticulate>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H. (2022). *Stringr: Simple, consistent wrappers for common string operations*. <https://CRAN.R-project.org/package=stringr>
- Wickham, H. (2023). *Forcats: Tools for working with categorical variables (factors)*. <https://CRAN.R-project.org/package=forcats>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., & Bryan, J. (2023). *Readxl: Read excel files*. <https://CRAN.R-project.org/package=readxl>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. <https://CRAN.R-project.org/package=dplyr>
- Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. <https://CRAN.R-project.org/package=purrr>
- Wickham, H., Hester, J., & Bryan, J. (2023). *Readr: Read rectangular text data*. <https://CRAN.R-project.org/package=readr>
- Wickham, H., & Seidel, D. (2022). *Scales: Scale functions for visualization*. <https://CRAN.R-project.org/package=scales>
- Wickham, H., Vaughan, D., & Girlich, M. (2023). *Tidyr: Tidy messy data*. <https://CRAN.R-project.org/package=tidyr>
- William Revelle. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Wright, B. D. (1977). Solving measurement problems with the rasch model. *Journal of Educational Measurement*, 97–116.

##NOTES ##Bias might suggest that rescaled a parameters are systematically larger than z under certain simulations (or not) Variance estimates might suggest that the standard error of rescaled values is larger than those values estimated directly (or not). If differences do exist, one could then go on to articulate the conditions under which they exist (i.e., high difficulty, low difficulty, non-normal distributions of the underlying trait), etc...

metrics Because of simulation data with consistent underprediction, modifications were applied to both indices. A slight alteration to this index was made in the current inves-

tigation whereby the simpler direct inverse of the standard normal deviate was retained.

Note Find citations about extreme b parameters due to poor fit.