Item Characteristic Curve specification from Classical Test Theory descriptive indices

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

[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 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 coefficient 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 low within this simulated dataset ($\overline{DIF}$ = .08 on our 13x1 dimensional plotting space). Applying the CTT-derived ICCs to two different applied samples of 10,000 real life test takers resulted in a comparable mean DIF estimate of .12. An `R` package, `ctticc`, performs the ICC calculations presented in the current paper and provides test specialists with visual representations of CTT-derived item characteristics. ExternalDataRequests@ETS.org Laura Ballard-Todd: lballard@ets.org, Jonathan Steinberg: jsteinberg@ets.org

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 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 approaches estimate the parameters 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

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

[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 here is commonly denoted as "p(1)" or "p(1.0)".

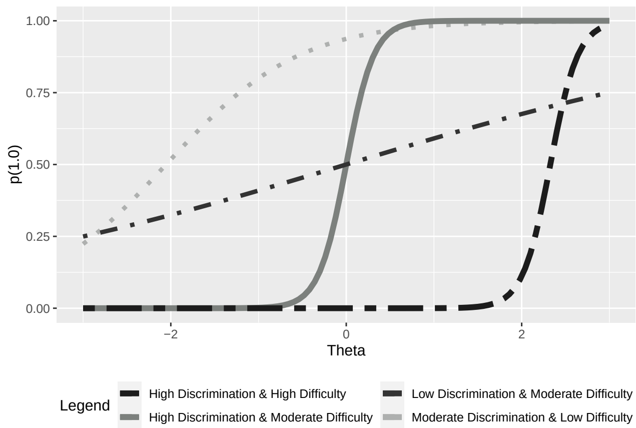

Figure 1. Item characteristic curves demonstrating differences in item difficulty and discrimination.

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 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 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 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, ICC's 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). 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 (e.g., Kulas et al., 2017).

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. 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. Within CTT examinations, for example, 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.

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).

Fan (1998) 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

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

[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 extent of the magnitude of this relationship.

provides enough empirical evidence on the extent of disparity between the two frameworks and the superiority of IRT over CTT despite the theoretical differences".

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 index[5]:

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

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 being operationalized as the *point-biserial* correlation between an item's binary response and the *corrected* total test score:

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

With g being 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. The current investigation retained the Kulas et al. (2017) formulas as "starting points" – because our intent was to generate ICCs comparable to IRT-derived ICCs, we also amended the Kulas et al. (2017) formulas to place z_g and $\hat{a}_i$ on metrics aligned with the IRT *b* and *a* parameters.

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 *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.

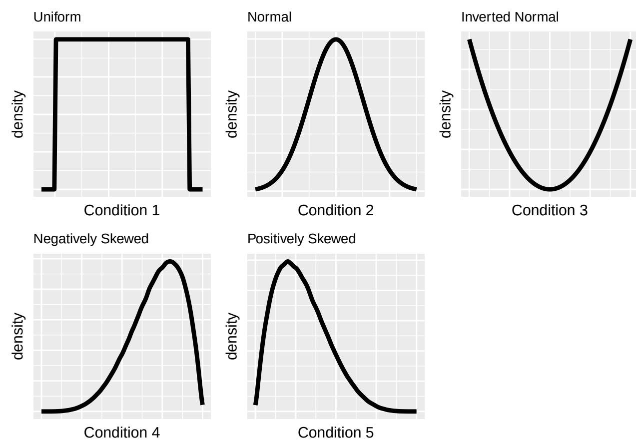

Figure 2. Shape of prescribed distributions of *p*-values across Study 1 conditions.

Study 1

Procedure and methods

We simulated datasets consisting of binary item responses. The simulated data prescriptively differed in distributions 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 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 2 provides a visual representation of idealized distributional forms that were prescribed across our simulations

Across all simulations, for items with extreme *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

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

responses of "1.0" were substituted. For items with p-values greater than 0.98, 200 random responses were given values of "0.0". This was done so the IRT models could converge without giving disproportionately extreme estimates.

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. Separate and overall regressions were applied to predict the IRT b parameter from the p-value derived z_g statistic to ensure that the relationship between b and z_g didn't depend on the distribution of item characteristics.

We used R (Version 4.2.3; R Core Team, 2023) and the R-packages *ape* (Version 5.7.1; Paradis & Schliep, 2019), *ctticc* (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; Grolemund & 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), *tinylabels* (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) for all analyses and manuscript development.

Results

Across all five conditions, simulated distributions exhibited 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), and the average $\hat{a}_i$ was 1.46 (sd=0.43). A regression predicting b from Z_g showed a R^2=0.96. A moderated regression across all five conditions showed a ΔR^2=0.00, with an $F_{(1,4943349)}$=1,783.21.

The average difference between our a-estimate and pseudo-a statistic was 0.04, with a $t_{(4940181)}$=613.01, p<0.05. Visual inspection of this effect seem to occur such that a-parameters above 1 occasionally had consistently under-estimated $\hat{a}$ values.The average difference between our b-estimate and pseudo-b statistic was 0.00, with a $t_{(4940181)}$=-1.34, p=0.18.

Overall, the average slope was 1.54 and the average intercept of 0.00.

A modification of the Z_g was done such that the inverse of the standard normal deviate was retained instead of the absolute value. There was also a regression derived modifier place on the previous $\hat{a}$ estimated using Kulas et al. (2017). This was a value of 1.72. These modifications were both made to more closely approximate the IRT metric.

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

Procedure and Methods

The purpose of study 2 is to generate ICC's based on the IRT model and 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. We used a simulated dataset on Wingen and two real-world datasets. The real-world dataset represented responses from 10,000 Test of English as a Foreign Language (TOEFL) iBT test-takers. This test measures all four academic English skills: reading (k=39), listening (k=40), and speaking (k=35). The two TOEFL ITP datasets come from two test forms and include item-level scores. Each form includes data from 10,000 examinees, and the examinee populations for the two forms do not overlap. Although ogives could be specified directly from the CTT-derived statistics, we made a procedural decision to retain the IRT 2PL as our functional definition for both IRT and CTT ogive specification:

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

Results

The area between ICC's was calculated between CTT-derived and IRT-derived ICC's. The theta range was between -6 and 6, with a maximum possible 2-dimensional area of 13. The average difference for all 100 curves from the simulated Wingen data was 0.08[6]. As we can see in figure **??**, most of

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

the 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 0.12. Applying the CTT-derived ICC functions to the real-world TOEFL data and computing the DIF with the IRT-derived ICC function, the average difference for the 228 curves was 0.12.

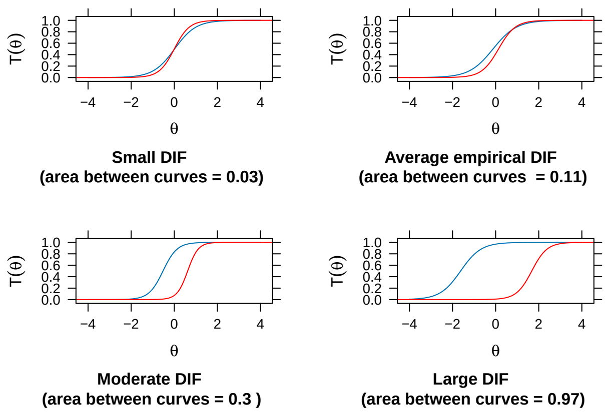

Figure 3. Four ICCs highlighting the difference between CTT and IRT-derivated ICCs at different levels of DIF.

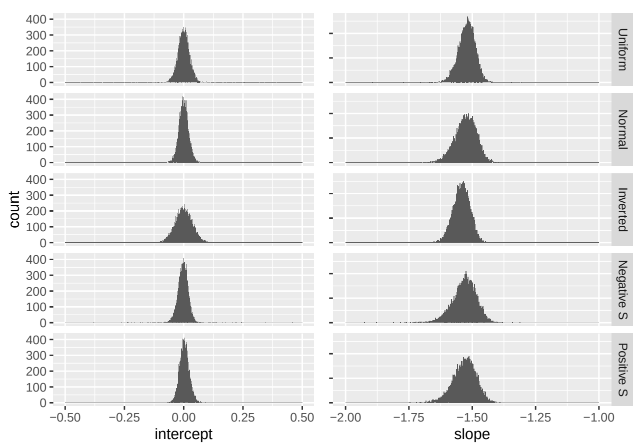

Figure 4. Individual intercepts and slopes grouped by study 2 simulation.

The average DIF across all 7 datasets was 0.11 (SD=0.14). There was no difference in average DIF across the 7 datasets ($F_{(6,321)}$ =1.84, p>.05). Out of the 328 items, only 17 (5.2%) had a DIF above 0.3.

The mirt package (Chalmers, 2012) was used to compute and plot the IRT statistics. To quantify the degree of difference between the two curves, the Area Between Curves was computed using Alfaro et al. (2009)'s package. Figure 3 presents some example ICCs exhibiting small, moderate, and relatively large levels of DIF. 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

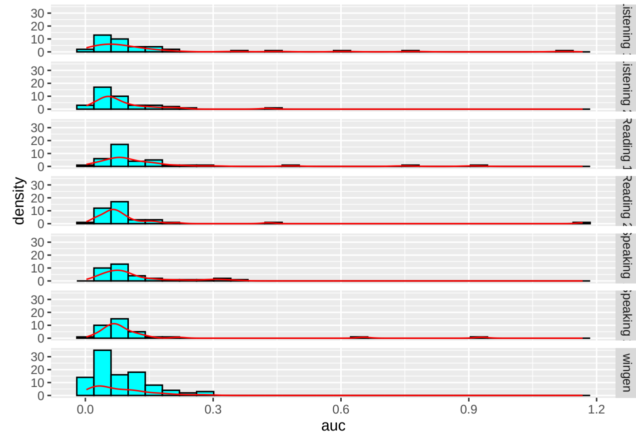

Figure 5. Histogram of all areas between ICCs plotted using IRT parameters vs ICCs plotted using CTT parameters.

item-total correlations, re-scaling and modifying them with Kulas et al. (2017) formulas).

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 when it comes to plotting visual representations such as ICC's. Practitioners and researchers that don't use IRT or Rasch models and

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 DIF. Base R seems to be the more accurate method.

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. 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.

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.

IRT analyses are also data hungry. These CTT-derived ICC estimates may be useful to individuals who wish to ultimately apply IRT, but are limited in… [maybe not]

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).

Visual inspection of this effect seem to occur such that 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.

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

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

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., & 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.

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.

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.

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

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., Grolemund, 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.

Appendix
Cut stuff

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".

##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. . . .

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])

2/9/2023 Notes: Check if the a parameter is estimated at the 0.5 location of the function. Research how the a parameter is scaled. Be more specific about the simulations. Write what we did when p-values were 0 and 1 for a column. Check the average a and b per simulation in line 255 For graph 7 update it by stacking the results we got from our simulations with the real data from ETS

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).

metrics Because of simulation data with consistent under-prediction, modifications were applied to both indices. A slight alteration to this index was made in the current investigation whereby the simpler direct inverse of the standard normal deviate was retained.

[^6] [^6]: We noted throughout our investigations that the "pseudo" *a* was systematically underpredicting the actual IRT a-parameter, so we ran regressions to further modify the "pseudo"-a scaling from the original Kulas et al. (2017) formula. Our regression modification added a further slope coefficient of 1.72 which resulted in a more precise rescaling of the CTT corrected item-total correlation.