# Item Characteristic Curves generated from common CTT Item Statistics

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

[1] Montclair State University

## Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. John T. Kulas: .

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

# Abstract

Item characteristic curves (ICCs) are visual indicators of important attributes of assessment items -*difficulty* and *discrimination*. Assessment specialists who examine Item Characteristic Curves usually do so from within the psychometric framework of either Item Response Theory or Rasch modeling.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 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 opposition for them not to do so, as ICCs based on CTT parameters could provide snapshot psychometric information as valuable as those gained from IRT- or Rasch-derived ICCs.Here we simulated data and plotted ICCs 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 ICCs. Practitioners and researchers that don't use IRT or Rasch models and instead opt to follow a CTT philosophy would benefit from having ICCs that use CTT statistics.

*Keywords:* keywords

Word count: X

# Item Characteristic Curves generated from common CTT Item Statistics

Item characteristic curves are frequently referenced by psychometricians as visual indicators of important attributes of assessment items - most frequently *difficulty* and *discrimination*. Within these visual presentations the x-axis ranges along "trait" levels (by convention annotated with the greek $\theta$), 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, and the y-axis probability refers to the likelihood of a "correct" response. From this visualization, the observer extracts the likelihood that respondents of any trait level would answer a focal 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 differently, it does not take much $\theta$ 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. If the curve is sharp (e.g., strongly vertical), this indicates high discrimination; if it is flatter, that is an indication of poorer discrimination.

Assessment specialists who examine ICCs usually do so from within the psychometric framework of either Item Response Theory or Rasch modeling. These frameworks provide the parameters necessary to plot the curves. Rasch models only estimate difficulty, and assume that differences in discrimination represent flaws in measurement. The IRT 2 parameter logistic model (2PL), however, models both item difficulty as well as item discrimination. Item difficulty (the $b$-parameter) is scaled as the trait level associated with a 50% likelihood of correct response (e.g., it is scaled to $\theta$). Item discrimination ($a$-parameter) is the degree to which an item differentiates across individuals who are characterized as being relatively lower or higher on the trait. 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 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 (to our knowledge!) attempt to represent them visually, as is common in IRT and Rasch applications. However, there is perhaps little opposition for them not to do so, as ICCs based on CTT parameters could provide snapshot psychometric information as valuable as those gained from IRT- or Rasch-derived ICCs. The largest obstacle to psychometricians regarding CTT-derived visuals is likely 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). 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) concluded that "no study provides enough empirical evidence on the extent of disparity between the two frameworks and the superiority of IRT over CTT despite the theoretical differences".

Fan (1998) looked at the correlations between ability estimates and item difficulty in CTT and all three IRT models. 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.[1]

Fan (1998) also investigated item invariance for all models. In theory, the major

[1] ...and in fact, as is presented below, the relationship between the IRT and CTT discrimination indices is non-linear - the correlation is an inappropriate index to capture the magnitude of this relationship.

advantage of IRT models over CTT is that the latter has an interdependency between the item and person statistics, whereas IRT has 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) found in his study, however, did not support the purported invariant advantage of IRT parameters over CTT indices. Both CTT item difficulty and discrimination degrees of invariance were highly correlated with those of IRT, indicating that they were highly comparable.

> NEED MORE - GRAB FROM OLD PAPER OR FIND NEW COMPARATIVE STUDIES

## Relationship between IRT and CTT indices

Lord (2012) described a function that approximates the relationship between IRT parameters and the CTT discrimination index of an item-test biserial correlation:

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

This formula wasn't intended for practical purposes but rather to assist in the conceptual comprehension of the discrimination parameter in IRT for people who were more familiar with CTT procedures. 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 (both simulated and utilizing real-world test data) identified systematic predictive differences across items with differing item difficulty values, so their recommended formula included a specification for item difficulty. This revised formula is used in the current presentation:

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

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 $g$. The transformation of the standard p-value was recommended in order to scale this index along an interval-level metric more directly anaologous to the IRT $b$ parameter. Figure XX 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 so at high levels of discrimination.

As we can see, the higher the corrected item-total correlations, the higher the estimated IRT a-parameter (discrimination). Also, as the p-values (difficulty) deviates from 0, the relationship between the estimated IRT a-parameter and the corrected item-total correlations becomes stronger.

Practitioners and researchers that don’t use IRT or Rasch models and instead opt to follow a CTT philosophy would benefit from having ICCs that use CTT statistics. This study intends to show evidence of the overlapping nature of CTT and IRT parameters when it comes to plotting ICCs.

## Study 1 - Estimating CTT-Derived ICCs

The purpose of study 1 is to look at the visualizations resulting from Kulas et al. (2017) formula on simulated data. We hypothesize that the relationship between the estimated IRT a-parameter and the corrected item-total correlations will be stronger as the later deviates from 0, which would mean that the item has more discrimination.

### Procedure and methods

We simulated data using Han (2007) software. 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 was 2, and the standard deviation 0.8. The mean for parameter b was 0 and the standard deviation 0.5.

### Results

## Study 2 - Evaluating the Comparability of IRT and CTT ICCs

The purpose of study 2 is to simulates a lot of test data and then generate ICCs based on the IRT model and then we compare that to our CTT estimates.

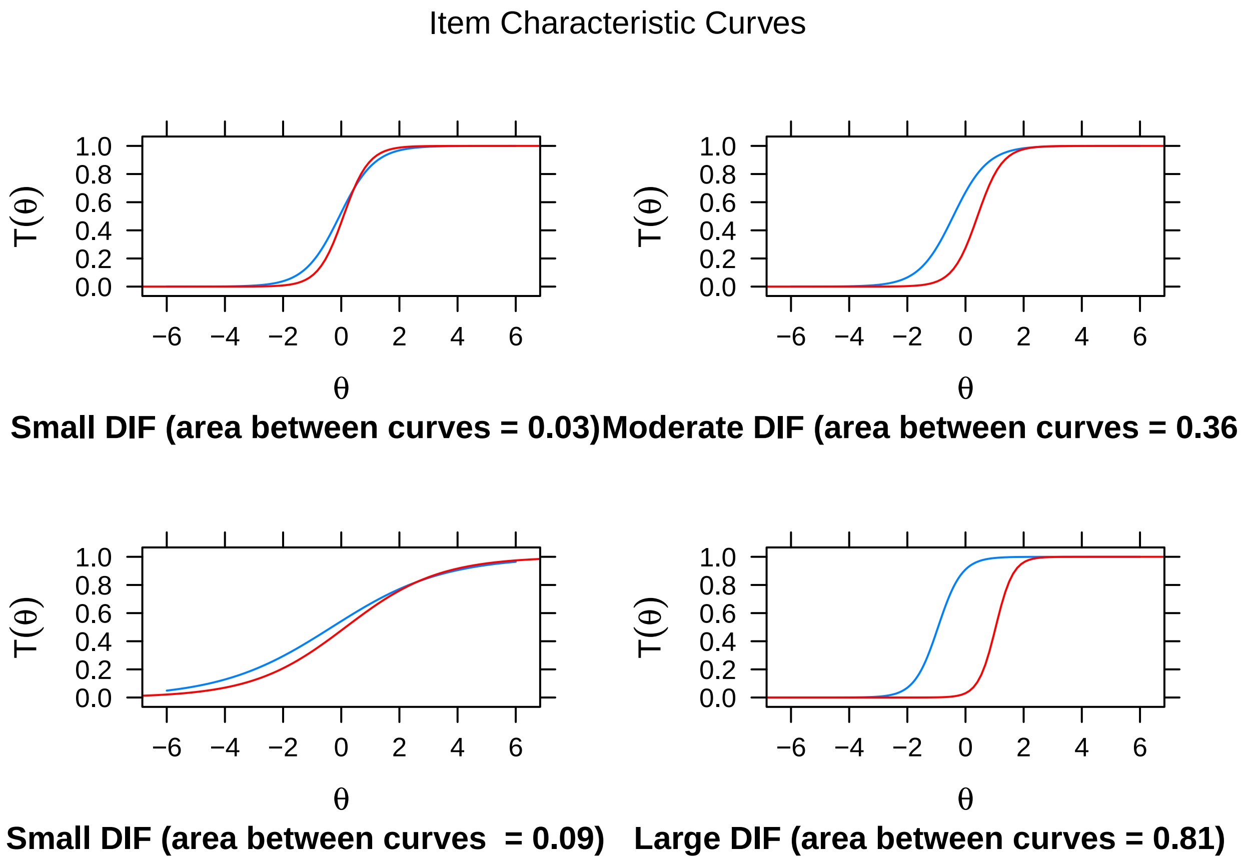
### Procedure and materials

The same simulated data as in study 1 was used. The mirt package was used to compute the IRT statistics. 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, modifying them with Kulas et al. (2017) formulas).

### Results

We used R (Version 4.0.3; R Core Team, 2020) and the R-packages *}ape* [@}R-ape], *dplyr* (Version 1.0.7; Wickham, François, et al., 2021), *DT* (Version 0.19; Xie et al., 2021), *forcats* (Version 0.5.1; Wickham, 2021a), *formattable* (Version 0.2.1; Ren & Russell, 2021), *geiger* (Version 2.0.7; Alfaro et al., 2009; Eastman et al., 2011; Harmon et al., 2008; Pennell et al., 2014; Slater et al., 2012), *ggplot2* (Version 3.3.5; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *irtplay* (Version 1.6.2; Lim, 2020), *jpeg* (Version 0.1.9;

Urbanek, 2021), *knitr* (Version 1.33; Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008; Sarkar & Andrews, 2019), *latticeExtra* (Version 0.6.29; Sarkar & Andrews, 2019), *markdown* (Version 1.1; Allaire et al., 2019; Xie et al., 2018, 2020), *mirt* (Version 1.34; Chalmers, 2012), *officer* (Version 0.3.19; Gohel, 2021), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *pdftools* (Version 3.0.1; Ooms, 2021), *psych* (Version 2.1.6; Revelle, 2021), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 2.0.1; Wickham & Hester, 2021), *readxl* (Version 1.3.1; Wickham & Bryan, 2019), *reticulate* (Version 1.20; Ushey et al., 2021), *rmarkdown* (Version 2.10; Xie et al., 2018, 2020), *shiny* (Version 1.6.0; Chang et al., 2021), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.1.4; Müller & Wickham, 2021), *tidyr* (Version 1.1.3; Wickham, 2021b), *tidyverse* (Version 1.3.1; Wickham, Averick, et al., 2019), and *tinytex* (Version 0.33; Xie, 2019) for all our analyses. The area between ICC's was calculated between CTT-derived and IRT-derived ICCs. The average difference for all 100 curves was 0.35.



Small DIF (area between curves = 0.03)

Moderate DIF (area between curves = 0.36

Small DIF (area between curves = 0.09)

Large DIF (area between curves = 0.81)

## Results

## Discussion

If this general idea is well-recieved (SIOP members would seem to represent a great barometer!) we would like to stress the CTT ICCs 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 extremly 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.

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

Alfaro, M., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D., Carnevale, G., & Harmon, L. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 13410–13414.

Allaire, J., Horner, J., Xie, Y., Marti, V., & Porte, N. (2019). *Markdown: Render markdown with the c library 'sundown'.* https://CRAN.R-project.org/package=markdown

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* https://github.com/crsh/papaja

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

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r.* https://CRAN.R-project.org/package=shiny

Eastman, J., Alfaro, M., Joyce, P., Hipp, A., & Harmon, L. (2011). A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, *65*, 3578–3589.

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.

Gohel, D. (2021). *Officer: Manipulation of microsoft word and powerpoint documents.*

https://CRAN.R-project.org/package=officer

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.

Han, K. (2007). WinGen3: Windows software that generates irt parameters and item responses [computer program]. *Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst*.

Harmon, L., Weir, J., Brock, C., Glor, R., & Challenger, W. (2008). GEIGER: Investigating evolutionary radiations. *Bioinformatics*, *24*, 129–131.

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. https://CRAN.R-project.org/package=purrr

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.

Lim, H. (2020). *Irtplay: Unidimensional item response theory modeling*. https://CRAN.R-project.org/package=irtplay

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

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. https://CRAN.R-project.org/package=tibble

Ooms, J. (2021). *Pdftools: Text extraction, rendering and converting of pdf documents*. https://CRAN.R-project.org/package=pdftools

Pennell, M., Eastman, J., Slater, G., Brown, J., Uyeda, J., Fitzjohn, R., Alfaro, M., & Harmon, L. (2014). Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*, 2216–2218.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ren, K., & Russell, K. (2021). *Formattable: Create 'formattable' data structures*. https://CRAN.R-project.org/package=formattable

Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. Springer. http://lmdvr.r-forge.r-project.org

Sarkar, D., & Andrews, F. (2019). *LatticeExtra: Extra graphical utilities based on lattice*. https://CRAN.R-project.org/package=latticeExtra

Slater, G., Harmon, L., Wegmann, D., Joyce, P., Revell, L., & Alfaro, M. (2012). Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution*, *66*, 752–762.

Urbanek, S. (2021). *Jpeg: Read and write jpeg images*. https://CRAN.R-project.org/package=jpeg

Ushey, K., Allaire, J., & Tang, Y. (2021). *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. (2019). *Stringr: Simple, consistent wrappers for common string operations*. https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr

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. (2019). *Readxl: Read excel files.* https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation.* https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2021). *Readr: Read rectangular text data.* https://CRAN.R-project.org/package=readr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/

Xie, Y. (2019). TinyTeX: A lightweight, cross-platform, and easy-to-maintain latex distribution based on tex live. *TUGboat*, *1*, 30–32. http://tug.org/TUGboat/Contents/contents40-1.html

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide.* Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Cheng, J., & Tan, X. (2021). *DT: A wrapper of the javascript library 'datatables'.* https://CRAN.R-project.org/package=DT

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook.* Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook

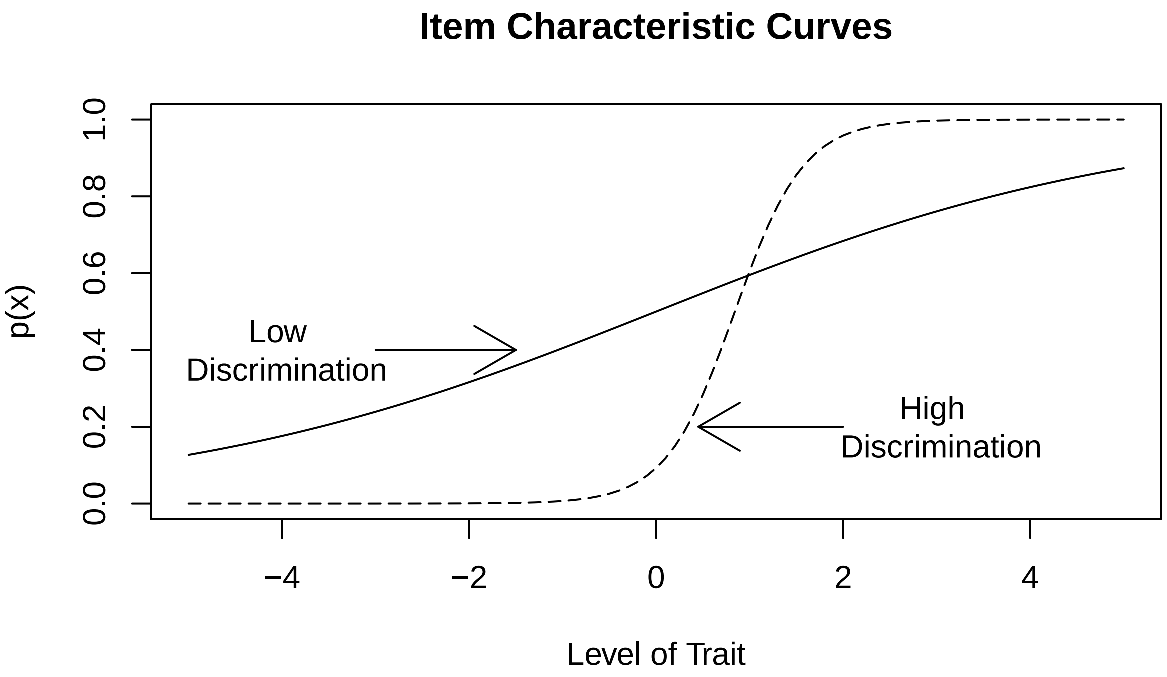


*Figure 1*. Item characteristic curves primarily reflecting differences in discrimination.

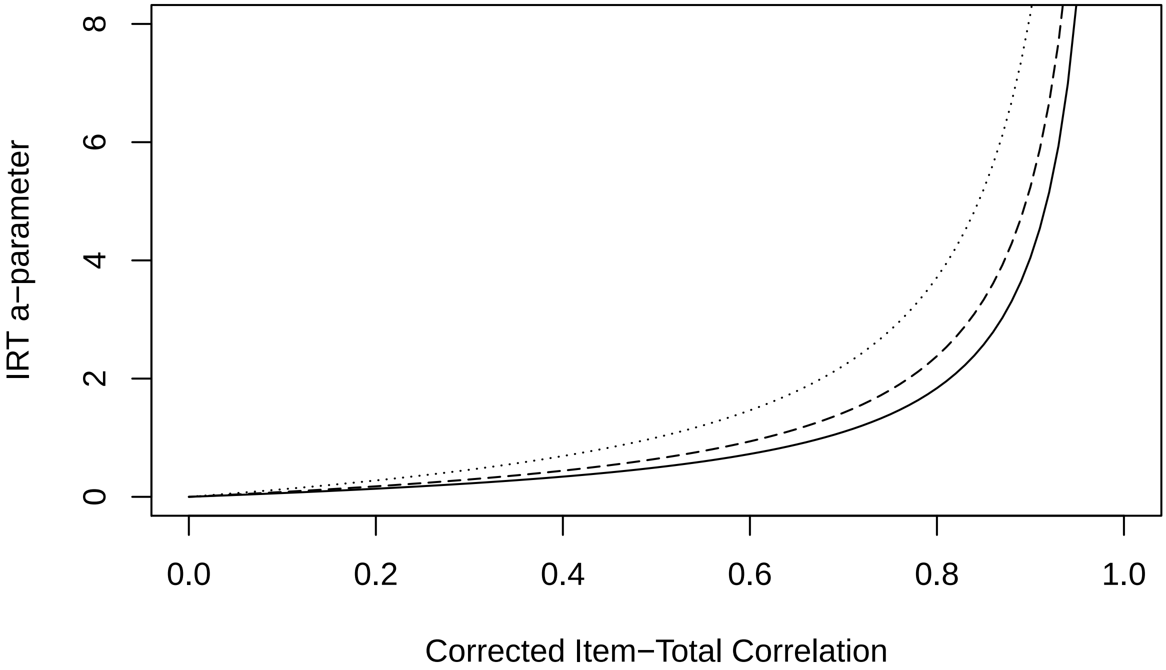


*Figure 2*. Empricially-derived functional relationship between the IRT *a* parameter and the CTT corrected-item total correlation as a function of item difficulty (p-value; solid = .5, dashed = .3/.7, dotted = .1/.9).

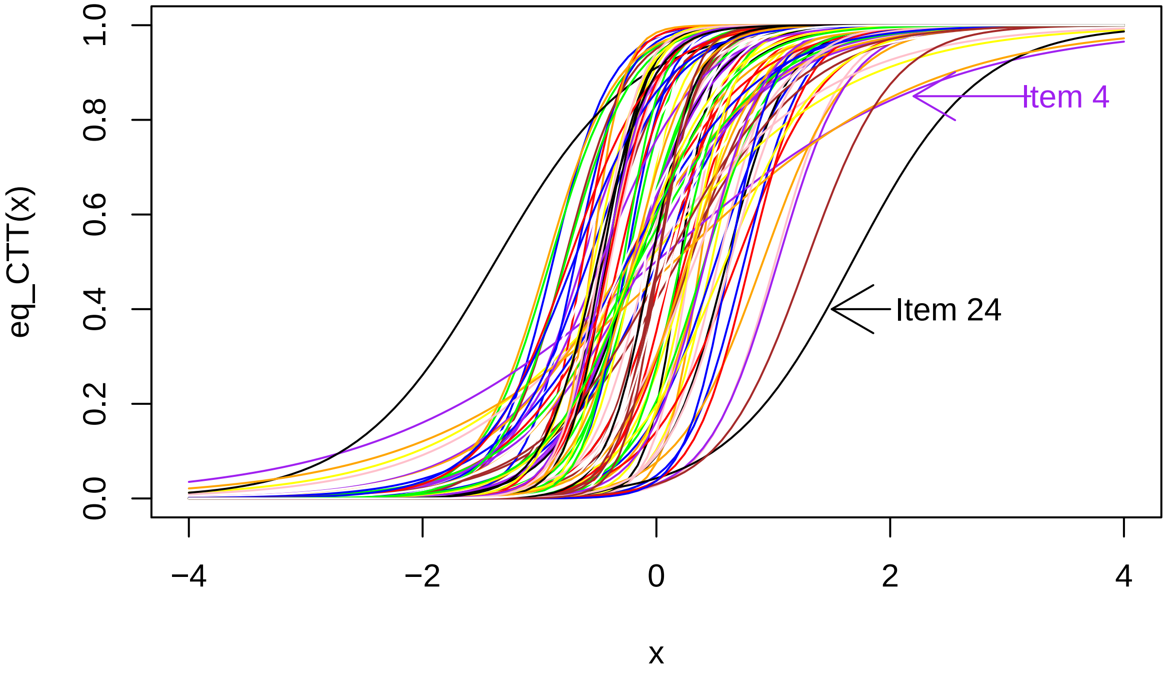


*Figure 3*. ICCs derived from only CTT parameters (with two noteworthy ICCs annotated).

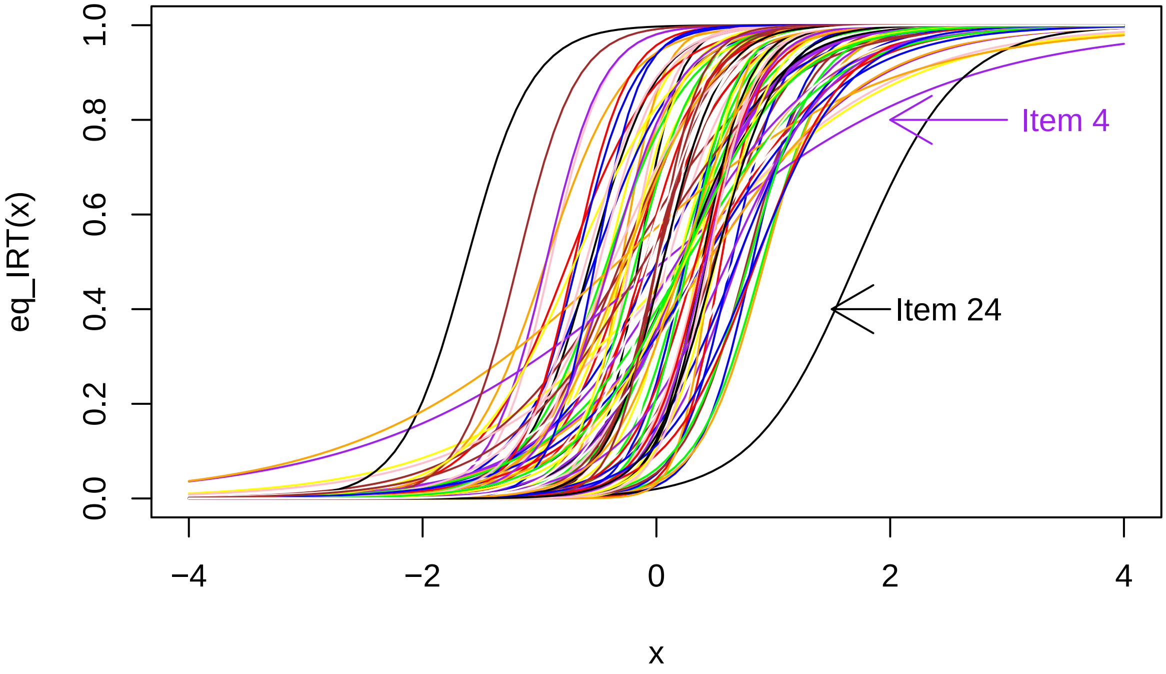


*Figure 4*. Typical ICCs derived from IRT parameters (same noteworthy items annotated).