Comparison of ICCs using IRT and CTT parameters

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.

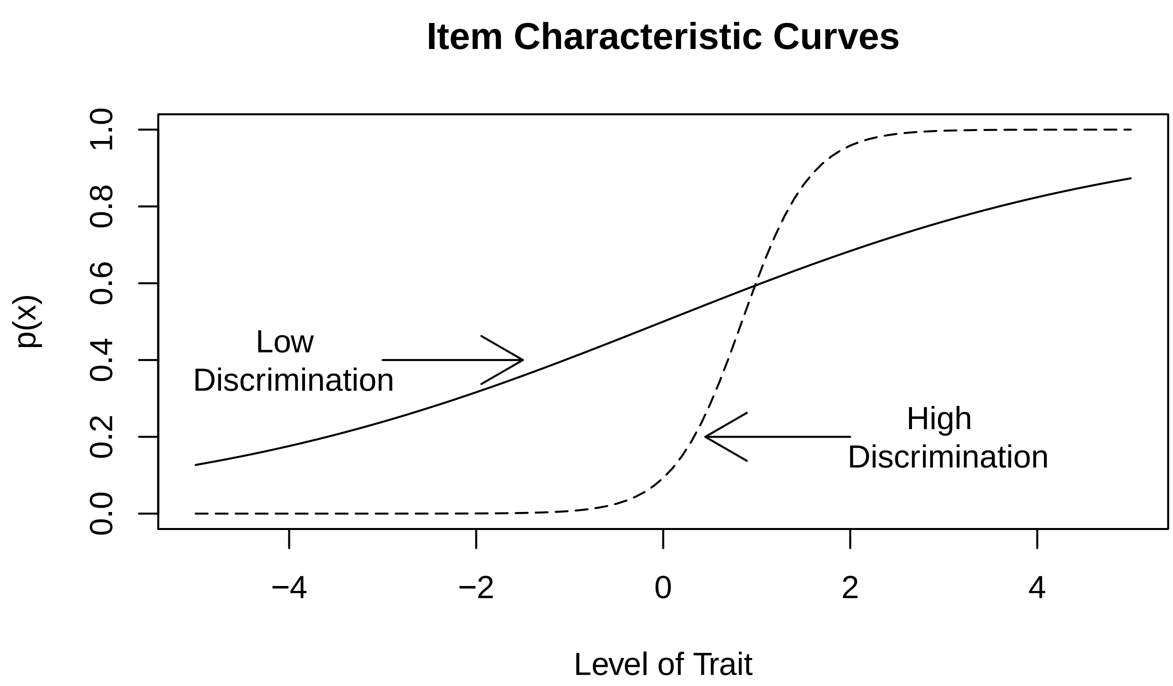
The authors made the following contributions. Diego Figueiras: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; John T. Kulas: Writing - Review & Editing.

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

Comparison of ICCs using IRT and CTT parameters

Introduction

Item characteristic curves are very often used by psychometricians to showcase and analyze the attributes of the item on a test or assessment. The x-axis shows a wide range of trait levels (ranging from high to low on the trait), while the y-axis displays probabilities of getting the item correct that range from 0 to 1. Each item has a curve. By looking at it, we can know the likelihood with which respondents of any trait level would answer any item correctly. If the curve is leaning towards the lower end of the trait level, this indicates that it is easy to answer the item correctly. On the contrary, if the curve is leaning towards the higher end of the trait level, this indicates that the item is difficult. If the curve is steep, this indicates high discrimination among respondents; if it is flat, it indicates no discrimination.



Psychometricians who examine ICCs usually do it using Item Response Theory and Rasch models to get the parameters necessary to plot the curves. In a 2PL model, these would be item difficulty and item discrimination. Item difficulty is the necessary trait level for a respondent to have a 50/50 chance to answer the item correctly. Item discrimination is the degree to which an item can differentiate among individuals with low and high levels of the trait. From a Classical Test Theory (CTT) frame of thinking, the difficulty of an item is determined by looking at the p-values of the items, while discrimination is determined by checking the Cronbach alpha and the corrected item total correlations. Psychometricians who look at these CTT parameters don't typically use them to plot ICCs.There is no reason for them not to, since ICCs based on CTT parameters could provide information as valuable as those based on IRT or Rasch without the need of being familiar with these models and with how to compute the necessary estimates. Fan states in summary that IRT and CTT "… framework produce very similar item and person statistics" (p.379).

There is research that shows that there is little difference between the parameters of both frameworks.@hambleton1993comparison 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) conducted a study to empirically test the differences between the two frameworks. According to him, "The findings here simply show that the two measurement frameworks produced very similar item and person statistics both in terms of the comparability of item and person statistics between the two frameworks and in terms of the degree of invariance of item statistics from the two competing measurement frameworks." In his study, Fan (1998) looked at the correlations between ability estimates and item difficulty in CTT and all three IRT models. These correlations were very high, between high .80 and low .90. As of item discrimination, correlations were moderate to high, with only a few being very low.

He also looked at the item invariance for all models. In theory, the major advantage of IRT models over CTT is that the latter has a circular dependency between the item and person statistics, while IRT has no such dependency, which means that the item parameters don't depend on the sample and the person parameters don't depend on the set of items. This property of invariance is very important, since item estimates can be used regardless of the sample you are giving the test or assessment to. An item will always have the same level of difficulty regardless of who is responding, for example.

What Fan (1998) got on his study, however, shows empirical evidence against this supposed advantage of IRT against CTT. The CTT item difficulty and discrimination degrees of invariance were highly correlated with those of IRT, indicating that they were highly comparable.

Lord (2012) described a function that approximates the relationship between IRT and CTT discrimination parameters. Although this 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, the formula was later modified by Kulas, Smith, and Xu (2017), with the purpose of minimizing the average residual.The formula is the following: [INSERT R EXPONENTIAL FORMULA]

Where r is the biserial corrected item total correlation of the item. Simulations identified systematic slope and inflection differences across item with differing b values, so the formula was further changed to include the following modifiers:

[INSERT FINAL FORMULA] Where g is the absolute deviation from 50% responding an item correctly and 50% responding incorrectly, and it's computed like this: g=|p-0.5|. Zg is the standard normal deviation associated with g. If we visualize the results of these re-specifications of Lord's formula using p-values (difficulty) of .5, .3 (or .7), and .1 (or .9), and corrected item total correlations (discrimination) of .3, .7 and .1, respectively, we get the following:

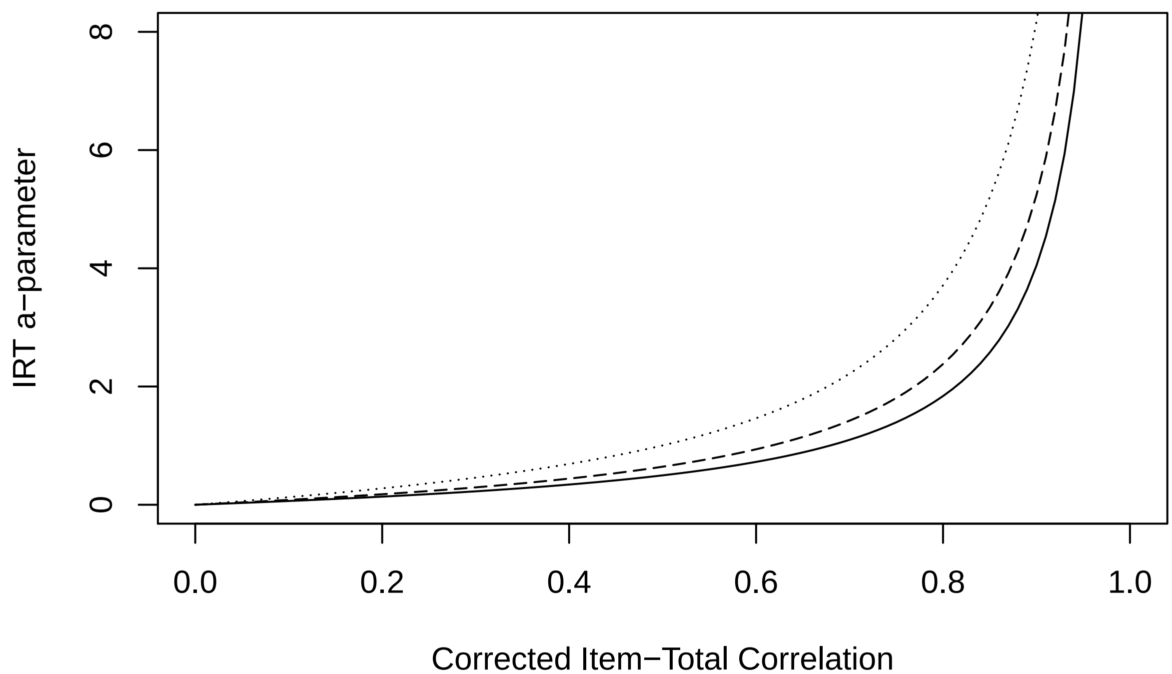


Figure 1. Functional relationship between the IRT *a* parameter and the CTT corrected-item total correlation as a function of item difficulty (p-value; solid = .5, dashed = .3/.7, dotted = .1/.9).

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.

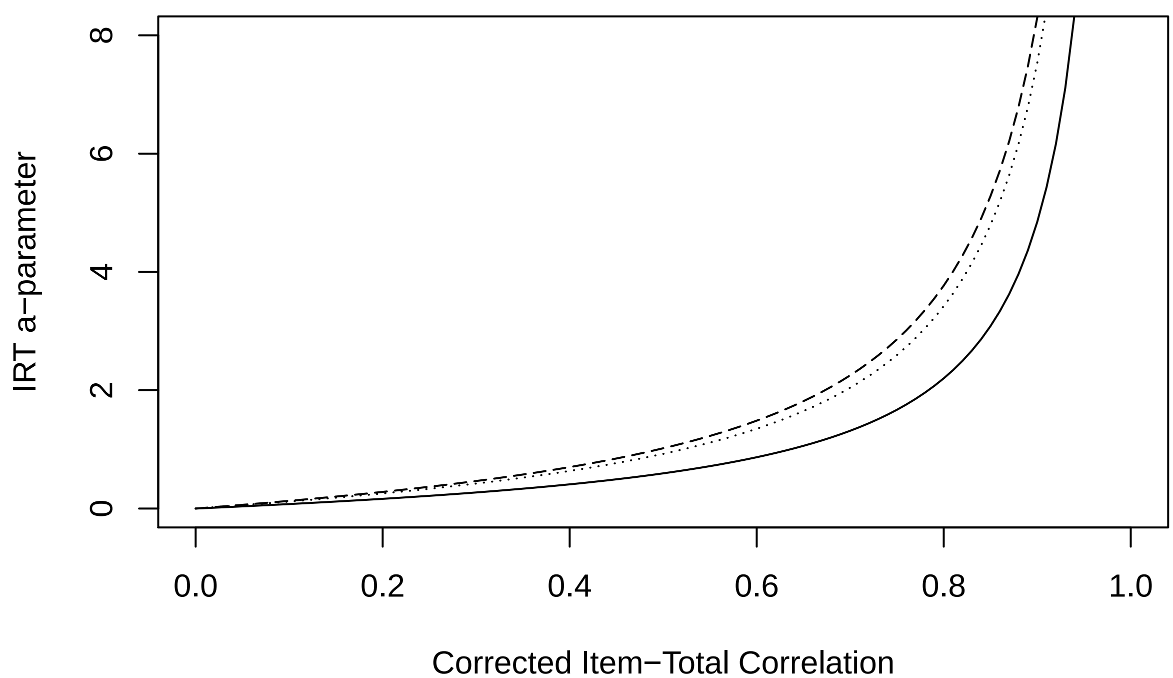
Study 1 - Visual of discrimination relationship

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 50, with response categories of either correct or incorrect (1 and 0).

Results



Study 2 - Item Characteristic Curves comparisons.

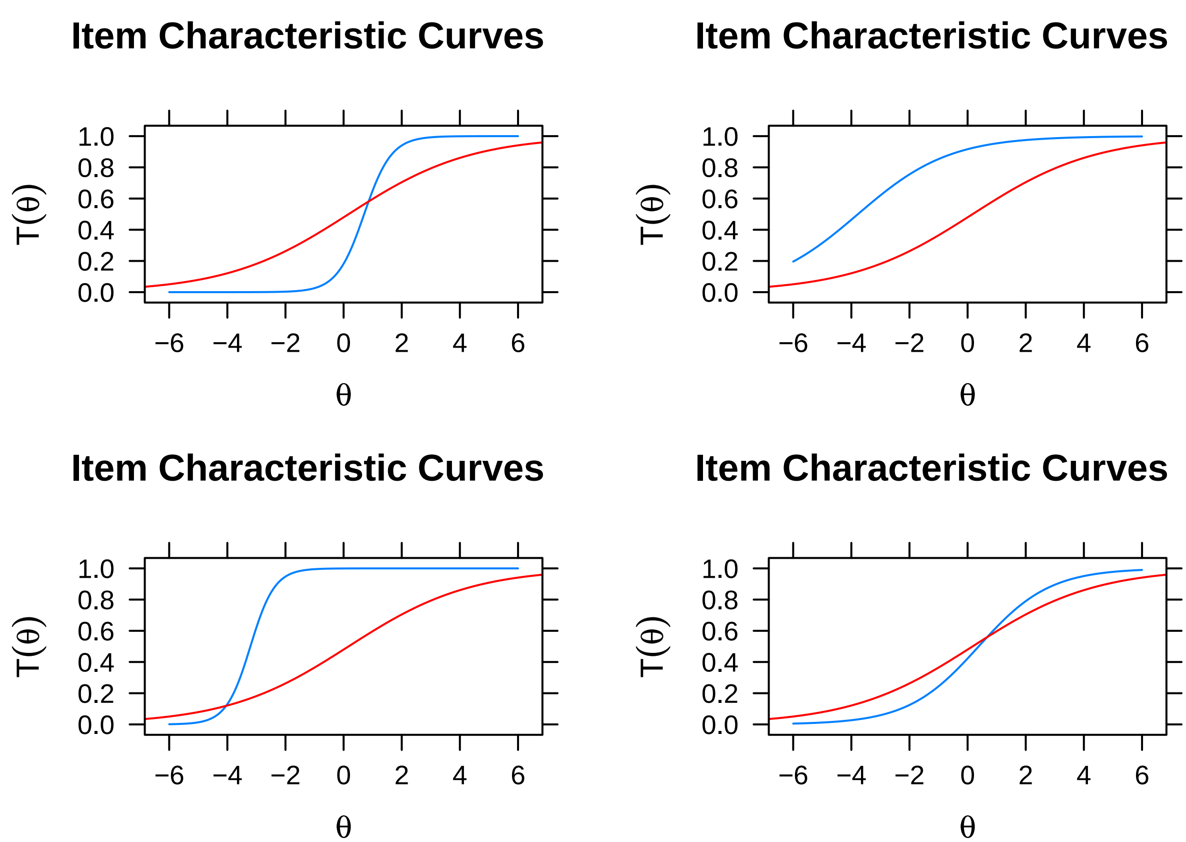
The purpose of study 2 is to simulates a lot of test data and then generate ICCs based on the IRT model and then we compare that to our CTT estimates.

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

```
## Iteration: 1, Log-Lik: -169092.337, Max-Change: 4.55861Iteration: 2, Log-Lik: -151096
```



Results

Discussion

References

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

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

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. (2012). *Applications of item response theory to practical testing problems*. Routledge.