1   Item Characteristic Curves derived from CTT Item Statistics

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

3   [1] Montclair State University

4   Author Note

¹¹   Item Characteristic Curves derived from CTT Item Statistics

¹²   **Introduction**

¹³   Item characteristic curves are very often used by psychometricians to showcase and

¹⁴ analyze the attributes of the item on a test or assessment. The x-axis shows a wide range

¹⁵ of trait levels (ranging from high to low on the trait), while the y-axis displays probabilities

¹⁶ of getting the item correct that range from 0 to 1. Each item has a curve. By looking at it,

¹⁷ we can know the likelihood with which respondents of any trait level would answer any

¹⁸ item correctly. If the curve is leaning towards the lower end of the trait level, this indicates

¹⁹ that it is easy to answer the item correctly. On the contrary, if the curve is leaning towards

²⁰ the higher end of the trait level, this indicates that the item is difficult. If the curve is

²¹ steep, this indicates high discrimination among respondents; if it is flat, it indicates no

²² discrimination.

²³   Psychometricians who examine ICCs usually do it using Item Response Theory and

²⁴ Rasch models to get the parameters necessary to plot the curves. In a 2PL model, these

²⁵ would be item difficulty and item discrimination. Item difficulty is the necessary trait level

²⁶ for a respondent to have a 50/50 chance to answer the item correctly. Item discrimination is

²⁷ the degree to which an item can differentiate among individuals with low and high levels of

²⁸ the trait. From a Classical Test Theory (CTT) frame of thinking, the difficulty of an item

²⁹ is determined by looking at the p-values of the items, while discrimination is determined by

³⁰ checking the Cronbach alpha and the corrected item total correlations. Psychometricians

³¹ who look at these CTT parameters don't typically use them to plot ICCs.There is no

³² reason for them not to, since ICCs based on CTT parameters could provide information as

³³ valuable as those based on IRT or Rasch without the need of being familiar with these

³⁴ models and with how to compute the necessary estimates. Fan states in summary that IRT

³⁵ and CTT "... framework produce very similar item and person statistics" (p.379).

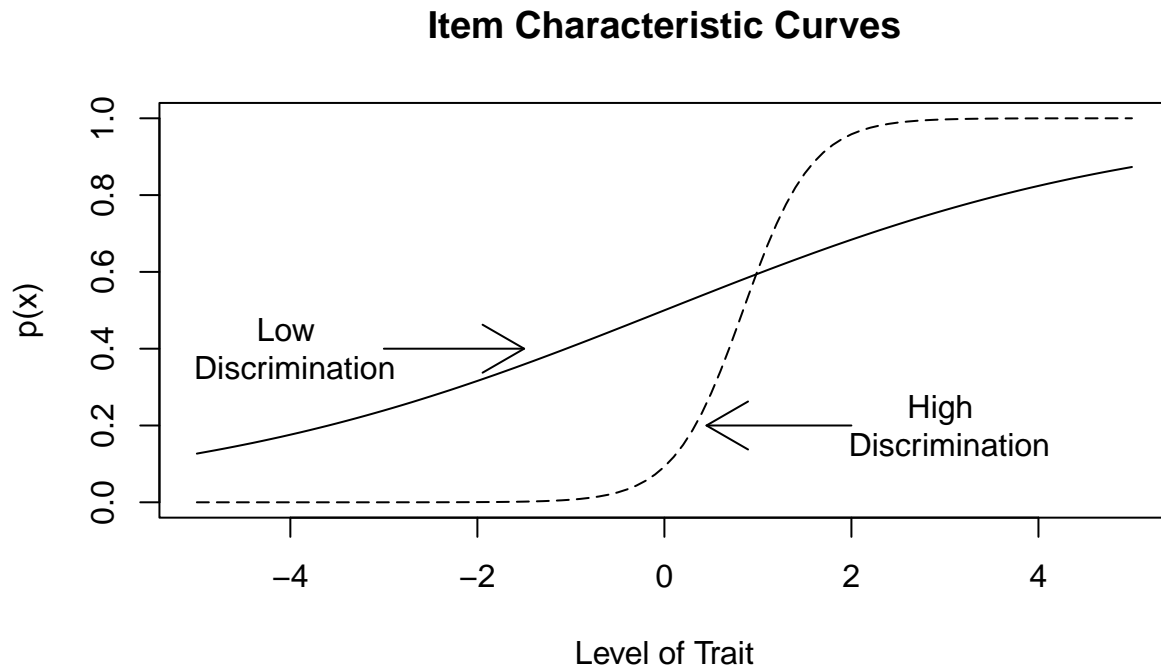³⁶   There is research that shows that there is little difference between the parameters of

Item Characteristic Curves

Figure 1. Item characteristic curve. X-axis indicates the trait level, and y-axis indicates the probability of getting the item correct.

37 both frameworks.@hambleton1993comparison concluded that "no study provides enough

38 empirical evidence on the extent of disparity between the two frameworks and the

39 superiority of IRT over CTT despite the theoretical differences."

40   Fan (1998) conducted a study to empirically test the differences between the two

41 frameworks. According to him, "The findings here simply show that the two measurement

42 frameworks produced very similar item and person statistics both in terms of the

43 comparability of item and person statistics between the two frameworks and in terms of

44 the degree of invariance of item statistics from the two competing measurement

45 frameworks." In his study, Fan (1998) looked at the correlations between ability estimates

46 and item difficulty in CTT and all three IRT models. These correlations were very high,

47 between high .80 and low .90. As of item discrimination, correlations were moderate to

48 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 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, Smith, and Xu (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).

Simulations identified systematic slope and inflection differences across item with differing item difficulty values, so the formula was further changed to include the following modifiers This revised formula is used in the current presentation:

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

71   Where g is the absolute deviation from 50% responding an item correctly and 50%

72 responding incorrectly (e.g., a "p-value" of .5). Z_g is the standard normal deviation

73 associated with g. The transformation of the standard p-value was recommended in order

74 to scale this index along an interval-level metric more directly anaologous to the IRT b

75 parameter. Figure XX visualizes the re-specifications of Lord's formula at p-values

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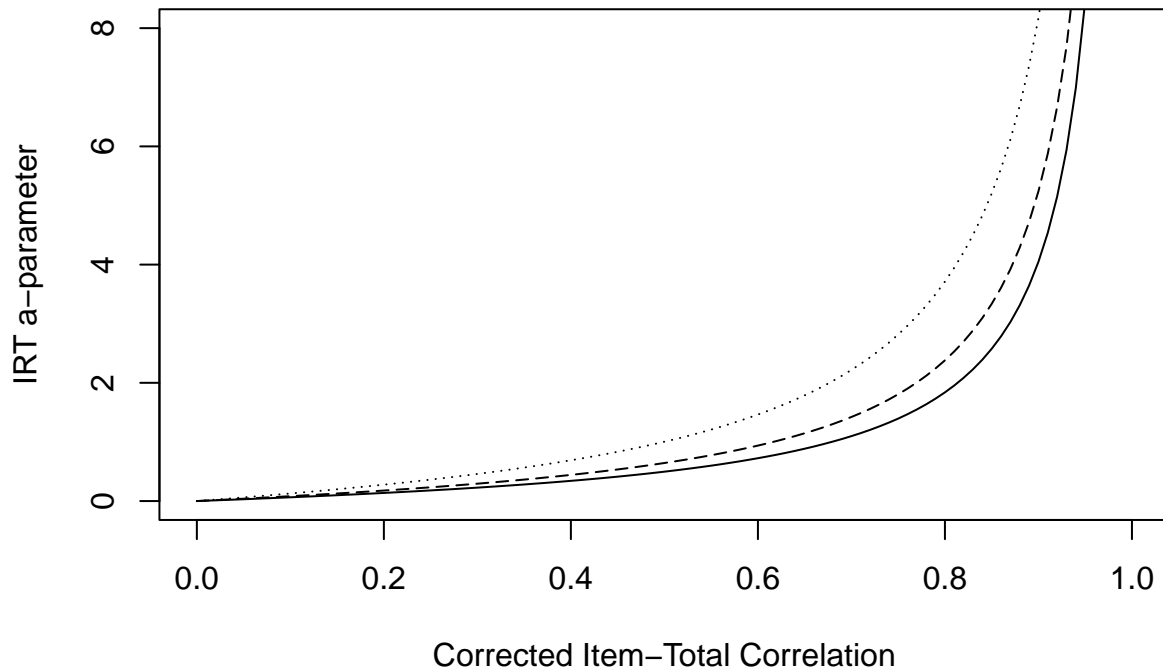
Figure 2. Functional relationship between the IRT a parameter and the CTT corrected-item total correlation as a function of item difficulty (p-value; solid = .5, dashed = .3/.7, dotted = .1/.9).

77   As we can see, the higher the corrected item-total correlations, the higher the

78 estimated IRT a-parameter (discrimination). Also, as the p-values (difficulty) deviates

79 from 0, the relationship between the estimated IRT a-parameter and the corrected

80 item-total correlations becomes stronger.

81   Practitioners and researchers that don't use IRT or Rasch models and instead opt to

82 follow a CTT philosophy would benefit from having ICCs that use CTT statistics. This

83 study intends to show evidence of the overlapping nature of CTT and IRT parameters

84 when it comes to plotting ICCs.

Study 1 - Visual of discrimination relationship

85

86   The purpose of study 1 is to look at the visualizations resulting from Kulas, Smith,

87 and Xu (2017) formula on simulated data. We hypothesize that the relationship between

88 the estimated IRT a-parameter and the corrected item-total correlations will be stronger as

89 the later deviates from 0, which would mean that the item has more discrimination.

Procedure and methods

90

91   We simulated data using Han (2007) software. Our sample was 10,000 observations,

92 with a mean of 0 and a standard deviation of 1. The number of items were 100, with

93 response categories of either correct or incorrect (1 and 0).The mean for the a parameter

94 was 2, and the standard deviation 0.8. The mean for parameter b was 0 and the standard

95 deviation 0.5.

96 **Results**

97
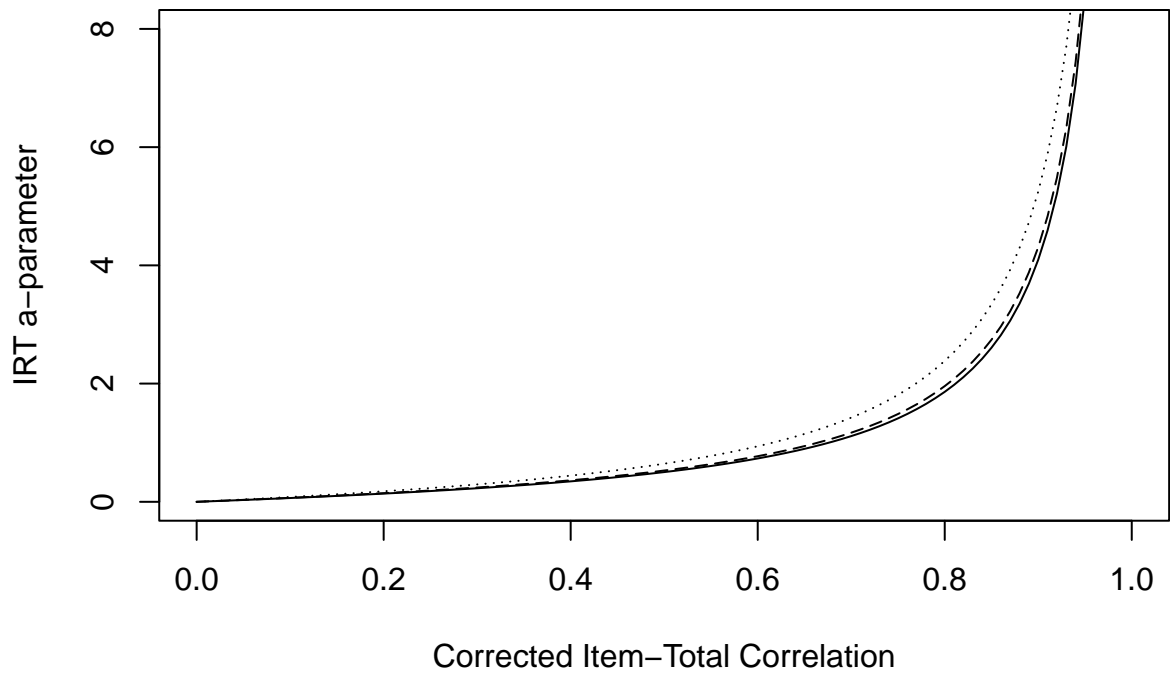
₉₈   **Study 2 - Item Characteristic Curves comparisons.**

₉₉   The purpose of study 2 is to simulates a lot of test data and then generate ICCs

₁₀₀ based on the IRT model and then we compare that to our CTT estimates.

₁₀₁ **Procedure and materials**

₁₀₂   The same simulated data as in study 1 was used. The mirt package was used to

₁₀₃ compute the IRT statistics. The blue curves were plotted using 2PL IRT parameters (a

₁₀₄ and b), while the red curves were plotted using CTT parameters (p-values and corrected

₁₀₅ item-total correlations, modifying them with Kulas, Smith, and Xu (2017) formulas).

₁₀₆ **Results**

₁₀₇   We calculated the area between the curves plotted using CTT parameters and the

₁₀₈ curves with IRT parameters. The average difference for all 100 curves was 0.35.

Item Characteristic Curves
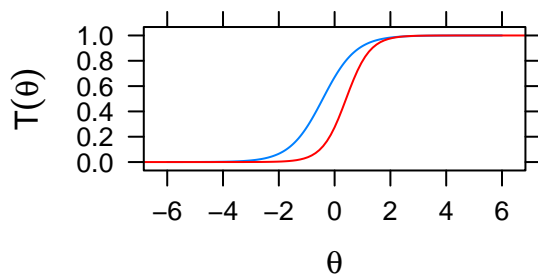
Medium area between curves (0.36)
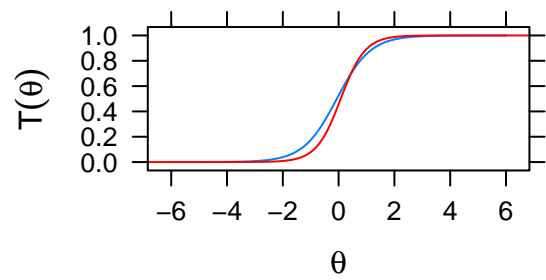
Low area between curves (0.03)
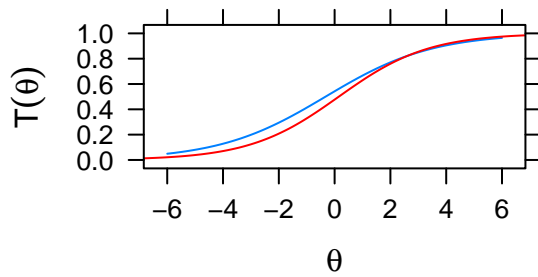
Low area between curves (0.09)
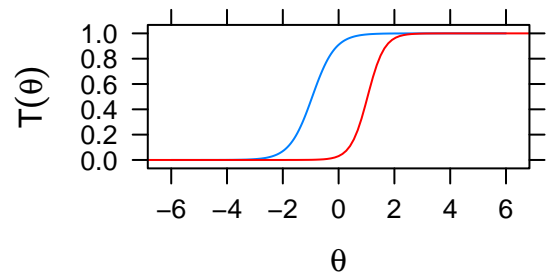
Big area between curves (0.81)

109

Results
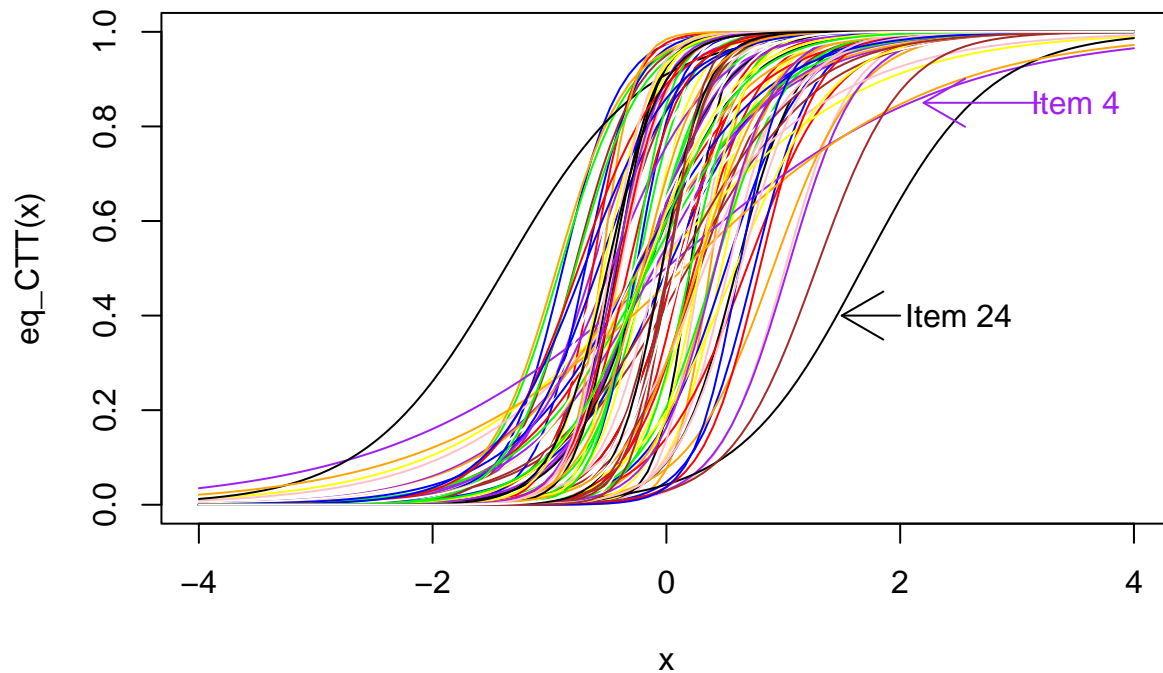
110

Discussion

111

Figure 3. ICCs using CTT parameters

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*
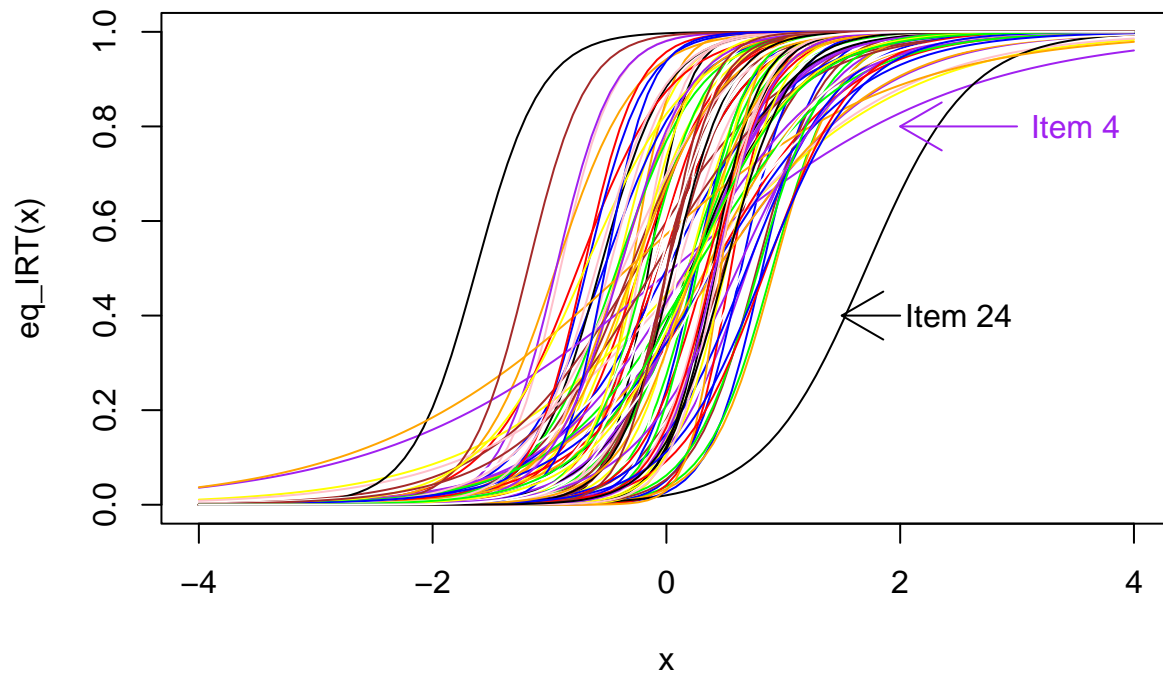
Figure 4. (#fig:irt curves)ICCs using IRT parameters

122   *Measurement, 18*(4), 393–407.

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

124   *problems*. Routledge.