Item Characteristic Curves generated from common CTT Item Statistics

Diego Figueiras[1] John Kulas[2]

[1] *Montclair State University*

[2] *eRg*

Introduction

Item characteristic curves (ICCs) are frequently referenced by psychometricians as visual indicators of important attributes of assessment items - most frequently `difficulty` and `discrimination`. This information is conveyed through ICCs (see Figure 1 for reference). Assessment specialists who examine ICC's usually do so from within the psychometric framework of either Item Response Theory (IRT) or Rasch modeling. We previously provided an extension of this tradition of item characteristic visualization within the more commonly leveraged Classical Test Theory (CTT) framework, but we did not focus on functional location. This current study builds on the first and focuses on placing the CTT p-value on the IRT b-parameter metric.

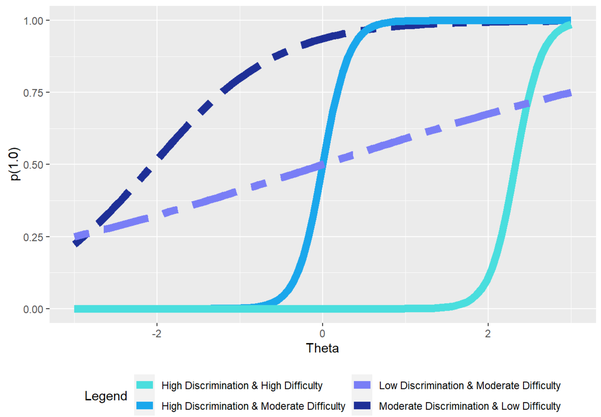


Figure 1: Item characteristic curves reflecting visual differences in difficulty and discrimination.

Method

We built five simulations of binary data, each with different distributions of p-values, as can be seen in Figure 2. Each simulation consisted of 10,000 observations and 100 items. Simulation 1 was uniform, with p-values ranging from 0 to 1. Simulation 2 was a normal distribution with p-values centered around 0.5. Simulation 3 was an inverted U-shaped distribution, with p-values ranging from 0 to 1. Simulation 4 was a left skewed distribution with p-values centered around 0.5, and simulation 5 was a right skewed distribution with p-values centered around 0.5.

We regressed *b*-parameters onto z_g indices of all simulations, and used the average slope and intercept to re-scale *p-values* on a pseudo-θ scale for graphing purposes.

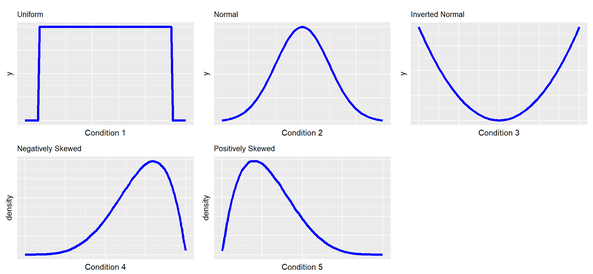
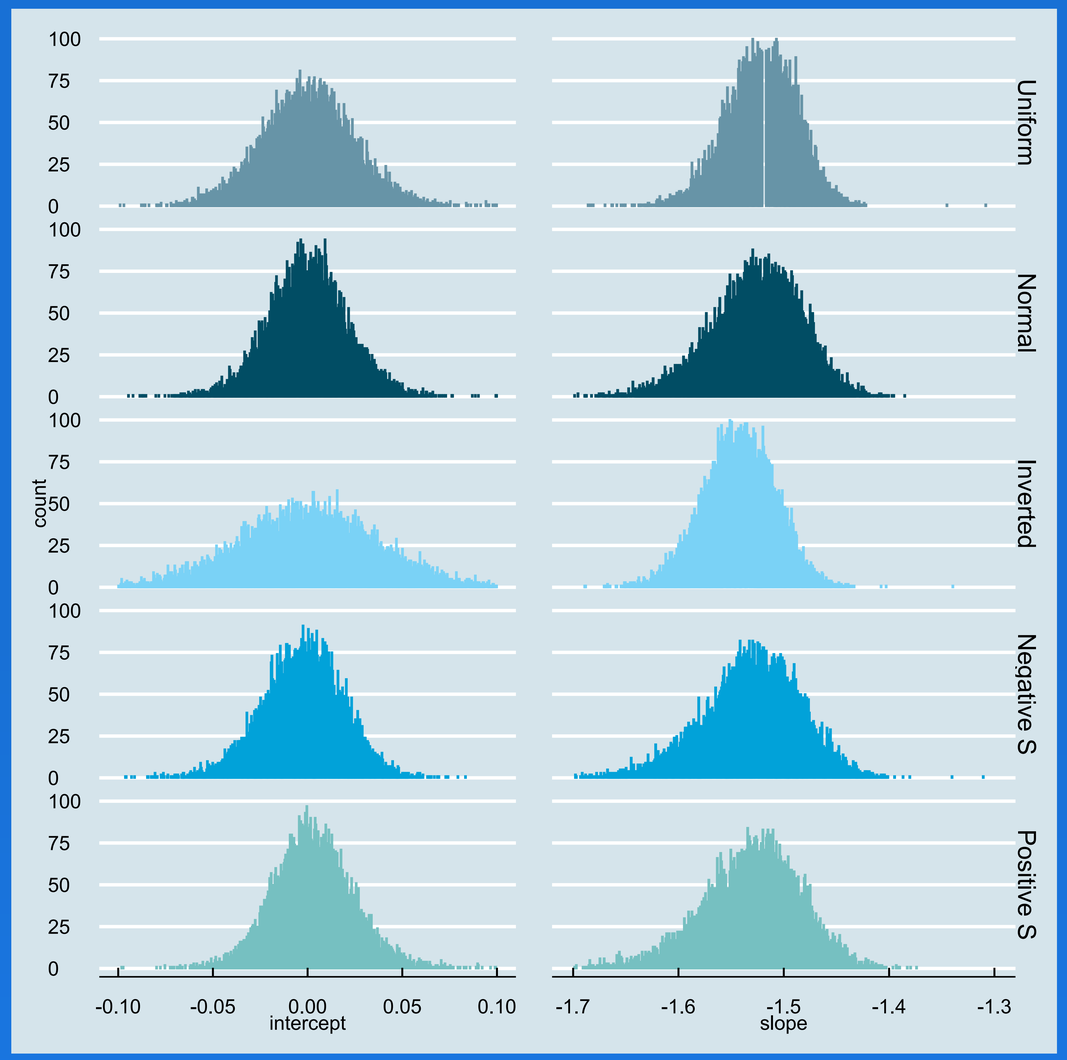


Figure 2: Shape of prescribed distributions of p-values across Study 1 conditions.

SIOP ANNUAL CONFERENCE

BOSTON and ONLINE • April 19-22, 2023



posterdown

Results

The resulting regression coefficients for all 5 simulations was an average intercept of approximately 0 and an average slope of -1.53. Two different one-way ANOVAs were applied with a non-significant mean intercept across conditions (F=0.66; p=0.62) and a statistically significant but meaningless mean slope effect (F=5.34; p=0, $\eta^2 = 0.0003$), as can be seen in the central figure. In this graph we present the empirical distributions of both slopes and intercepts for all 10,000 simulations per simulation condition. They are all centered at about -1.53, with very little deviance in terms of shape, kurtosis, or spread.[1].

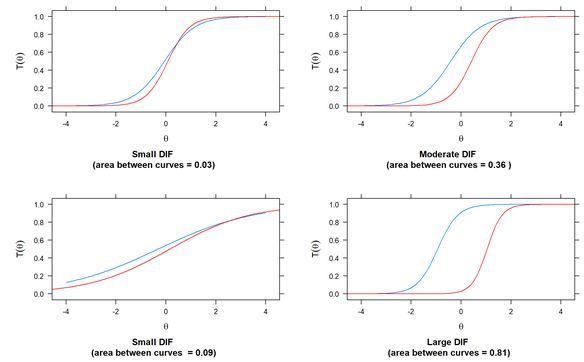


Figure 3: Four ICCs highlighting the difference between CTT and IRT-derivated ICCs at different levels of DIF.

The area between ICC's was calculated between CTT-derived and IRT-derived ICC's. The average difference for all 100 curves was 0.214. As we can see in Figure 5, most of the data is located at the lower end, indicating that out of the 100 items, most of them have areas between the curves of less than 0.21. This DIF was computed after scaling our Z_g using the coefficients estimated with our simulations. Without the regression coefficient modifier the average area under the curves was 0.80, as we can see in Figure 4. We ran a test of significance between these two means. Our results are t(99) = 11.72, p < .001.

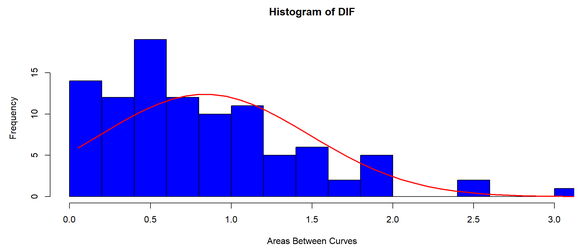


Figure 4: Histogram of DIF between ICCs plotted using IRT parameters vs ICCs plotted using CTT parameters.

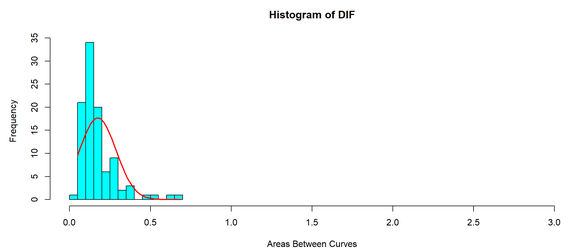


Figure 5: Histogram of DIF between ICCs plotted using IRT-paramets VS ICCs plotted using CTT statistics using regression coeficients modifier

Discussion

Although invariance is a property historically associated exclusively for IRT applications, large scale data, truly random sampling, and large range items can also yield stable CTT item and person statistics **(Fan, 1998; Kulas et al., 2017)**. The current investigation scaled the CTT p-value to the IRT b-parameter. The linking equation was relatively invariant across simulations. Our most relevant finding was that there was no interaction effect between simulated conditions. We are currently working on using real world data to cross-validate the findings from these simulations. We created an `R` package that generates these CTT-derived ICCs (https://github.com/MontclairML/ctticc). This poster was crafted via `posterdown` **(Thorne, 2019)**.

1. There were 3383 cases removed from the overall 500000 simulated items due to extreme b-estimates.↩