Informative vs uninformative prior distributions with characteristic curve linking methods

Abstract

Linking of two forms is an important task when using item response theory, particularly when two forms are administered to nonequivalent groups. When linking with characteristic curve methods, the ability distribution and weights associated with that distribution can be used to weight observations differently. These are commonly specified as equally spaced intervals from -4 to 4, but other options or distributional forms can be specified. The use of these different distributions and weights of the ability distributions will be explored with a monte carlo simulation. Primary simulation conditions will include sample size, number of items, number of common items, ability distribution, and randomly varying population transformation constants. Study results show that the linking weights have little impact on the estimation of the linking constants, however the underlying ability distribution of examinees does have significant impact. Implications for applied researchers will be discussed.

Item response theory (IRT) has become a popular statistical technique to estimate an individual's ability and the characteristics of items given an individual's response string; a process often called calibration. A commonly used IRT model for calibration is Birnbaum's three parameter logistic model (Birnbaum, 1968):

$$p(x_j = 1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{-Da_j(\theta - b_j)}}.$$

The equation above models the likelihood of a respondent correctly answering item j given the respondent's ability (θ), the item's difficulty (b_j), the item's discrimination (a_j), the pseudo-guessing parameter of the item (c_j), and a scaling constant D, which is commonly set to 1.702 to transform the logistic ogive into a normal ogive (Camilli, 1994; De Ayala, 2013). The three item parameters (a_j, b_j, c_j) make up the item response function which is a mathematical representation of the likelihood a respondent answers a given item correctly given the ability of the respondent. Simpler models can be obtained by fixing parameters to zero or one. A more thorough introduction to IRT can be found in De Ayala (2013) and Lord (1980).

With IRT, the ability scale is arbitrarily decided based on a linear transformation (De Ayala, 2013; Kolen & Brennan, 2004). It is common when calibrating to set the ability scale to have a mean of 0 and a standard deviation of 1. However, as a result of this arbitrarily set ability scale, calibrations for two groups need not have the same ability scale. As such, linking the two together is common to ensure that the individual's ability and the item parameters are on the same scale for use on an assessment. Method moment and characteristic curve linking methods are available to link two assessments that were given to nonequivalent groups. The characteristic curve methods also allow users to specify proficiency points and weights that mimic the ability distribution of the individuals tested. Only a single study, (S. Kim & Kolen, 2007), have empirically studied the choice of the proficiency points and weights with the characteristic curve methods.

This study aims to extend the work done by S. Kim and Kolen (2007) by empirically studying the impact of different specification of proficiency points and weights with the characteristic curve methods. In an attempt to generalize to many more empirical data conditions, a wide variety of simulation conditions are included in the design, including the transformation constant values, the number of common items, sample size, and length of assessment. The ability to accurately estimate the linking parameters will be the outcome of interest in this study.

Linking Methods

Two forms can be linearly linked together based on the following three equations (Kolen & Brennan, 2004):

$$\begin{aligned} \theta_o &= A\theta_n + B, \\ a_{jo} &= \frac{a_{jn}}{A}, \\ b_{jo} &= Ab_{jn} + B. \end{aligned} \tag{1}$$

In the first equation above, A and B represent the linear transformation constants that link the old and the new ability scales together, represented as θ_o and θ_n respectively. In order to keep the probability of answering the item correctly at a given ability the same after transforming the ability scale, the item parameters are adjusted based on the second two equations above. Finally, a_{jo}, a_{jn}, b_{jo}, and b_{jn} represent the discrimination and difficulty parameters for item j and for either the old form (o) or new form (n). Note: The pseudo-guessing parameter remains unchanged in the old and new form (i.e. $c_{jo} = c_{jn}$). The A and B transformation constants shown in Equation (1) are estimated from various linking methods discussed in more detail below.

A common linking design used in practice is the common-item nonequivalent group design (Kolen & Brennan, 2004) or sometimes referred to as an anchor instrument design (Holland & Dorans, 2006). In this design, two forms (called old and new) are given to two different groups of respondents. The responses from these two forms are then calibrated separately to place them each on their own "0,1" ability scale. Then, through the use of a set of common items that were taken by both groups of respondents, the scores and ability of the respondents can be linked together through linking methods. The main assumption made with these methods are that the two ability scales are linearly related such that the two forms can be adjusted using Equation (1). The nonequivalent group design is popular in that respondents do not need to be statistically equivalent when given the assessment, instead, they just need to be sampled from the same population. The goal is then to link or equate the two groups after responses are obtained.

Linking is not limited to only two groups, however when linking with more than two groups (e.g. grade level of individuals), the linking is done in a pairwise fashion with adjacent groups (Kolen & Brennan, 2004). Therefore, to keep the scope simpler, the discussion in this paper is limited to only two groups, but these methods naturally extend to more than two groups (as shown in the applied example).

There are four common linking methods used with the common item nonequivalent group design, mean/mean, mean/sigma, Haebara (Haebara, 1980), and Stocking-Lord (Kolen & Brennan, 2004; Stocking & Lord, 1983). A fifth method, concurrent calibration, is also possible to link the individual ability and item parameters when groups are nonequivalent. Concurrent calibration statistically adjusts the scales in a single calibration run again through the use of common items by estimating and adjusting the average ability for each group. In this method, it is common for the base group to have a "0,1" ability scale, however, the other groups will have different means and standard deviations to ensure the groups are equivalent. Concurrent calibration in simulation studies has been shown to provide better estimates than linking methods (S. Kim & Kolen, 2007).

Moment Methods

The mean/mean and mean/sigma methods, commonly called moment methods, use the mean or standard deviation of the discrimination or difficulty item parameters from the new and old forms to link them together. More specifically, the A and B linking constants for the mean/mean method can be calculated as:

$$\begin{aligned}\hat{A}_{mm} &= \bar{a}_{jo}/\bar{a}_{jn} \\ \hat{B}_{mm} &= \bar{b}_{jo} - \hat{A}_{mm}\bar{b}_{jn}\end{aligned} \tag{2}$$

where $\hat{A}_{mm}$ and $\hat{B}_{mm}$ are the estimated transformation constants for the mean/mean method using estimated item parameters. The estimated A parameter is computed by dividing the average discrimination of the old form by the average discrimination of the new form. The estimated B parameter is a function of the average difficulty parameters for the old and new forms, while taking into account the estimated A parameter.

The mean/sigma method is similar to the mean/mean method, but instead of using the average discrimination parameter to estimate the A parameter, a ratio of the standard deviations of the estimated difficulty parameters are used instead. This takes the following form:

$$\begin{aligned}\hat{A}_{ms} &= SD(\hat{b}_{jo})/SD(\hat{b}_{jn}) \\ \hat{B}_{ms} &= \bar{b}_{jo} - \hat{A}_{mm}\bar{b}_{jn}.\end{aligned} \tag{3}$$

Although simple, it has been shown that these methods do not perform quite as well as the other two methods, commonly referred to as the characteristic curve methods (Hanson & Béguin, 2002; S. Kim & Lee, 2006). In addition, although not explored in this study, mixed format assessments can be problematic for the moment methods (S. Kim & Lee, 2006). The interested reader is directed to J. Kim and Hanson (2002) and S. Kim and Lee (2006) for more information on using these methods with mixed format assessments.

Characteristic Curve Methods

The characteristic curve methods aim to minimize differences between the item characteristic curve or test characteristic curve for the Haebara (Haebara, 1980) and Stocking-Lord (Stocking & Lord, 1983) methods respectively. These procedures are computationally intensive and require numeric integration methods to approximate the minimization function (S. Kim & Lee, 2006; Kolen & Brennan, 2004).

More specifically, the Haebara (Haebara, 1980) method minimizes the following function:

$$H = H_1 + H_2, \tag{4}$$

where

$$H_1 = \sum_{ci}^{CI} \left[p(\theta_{jo}|\hat{a}_{jo}, \hat{b}_{jo}, \hat{c}_{jo}) - p(\theta_{jo}|\frac{\hat{a}_{jn}}{A}, A\hat{b}_{jn} + B, \hat{c}_{jn}) \right]^2 w_1(\theta_{jo}) \tag{5}$$

and

$$H_2 = \sum_{ci}^{CI} \left[p(\theta_{jn}|\hat{a}_{jn}, \hat{b}_{jn}, \hat{c}_{jn}) - p(\theta_{jn}|\hat{a}_{jo}A, \frac{\hat{b}_{jo} - B}{A}, \hat{c}_{jo}) \right]^2 w_2(\theta_{jn}), \tag{6}$$

where θ_{jo} and $w_1(\theta_{jo})$ are proficiency points and weights that represent the distribution of

θ_o (i.e. the old scale) and θ_{jn} and $w_2(\theta_{jn})$ are proficiency points and weights representing the distribution of θ_n (i.e. the new scale). To find the estimated A and B population linking constants, H is minimized using a computationally intensive search procedure that attempts to minimize differences in the common item characteristic curves.

The Stocking-Lord characteristic curve method is similar to the method by Haebara. With the Stocking-Lord approach, the following is minimized to find the estimated A and B parameters:

$$S = S_1 + S_2, \tag{7}$$

where

$$S_1 = \left[\sum_{ci}^{CV} p(\theta_{jo} | \hat{a}_{jo}, \hat{b}_{jo}, \hat{c}_{jo}) - p(\theta_{jo} | \frac{\hat{a}_{jn}}{A}, A\hat{b}_{jn} + B, \hat{c}_{jn}) \right]^2 w_1(\theta_{jo}) \tag{8}$$

and

$$S_2 = \left[\sum_{ci}^{CI} p(\theta_{jn} | \hat{a}_{jn}, \hat{b}_{jn}, \hat{c}_{jn}) - p(\theta_{jn} | \hat{a}_{jo}A, \frac{\hat{b}_{jo} - B}{A}, \hat{c}_{jo}) \right]^2 w_2(\theta_{jn}). \tag{9}$$

The proficiency points and weights are defined similarly as above from the Haebara method.

The primary difference between the Haebara and Stocking-Lord method can be seen in the equations with respect to when the summation is occurring. For the Haebara method, the summation is outside of the square brackets, this has the effect of the square differences is done at the item level. In comparison, the Stocking-Lord method sums over the common items prior to squaring. This is equivalent to finding the test characteristic curve for the common items. For this reason, small differences may be able to offset each other with the Stocking-Lord method. For example, a positive difference in one item may be offset by a negative difference on a second item.

Symmetric Linking

Symmetric linking takes into account information regarding the old and new scales, whereas unsymmetric linking only takes into account only one of the scales. For example with the Haebara method, symmetric linking would involve minimizing the difference with respect to both H_1 and H_2. Unsymmetric linking would only use either H_1 or H_2, but not both. Which function to minimize would change depending on which scale, old or new, is desired as the final scale, however traditionally when unsymmetric linking is used, only H_1 is used (S. Kim & Kolen, 2007). If the old scale is desired, H_1 would be used and if the new scale is desired, H_2 would be used. Similar explanations would be found for the Stocking-Lord methods where S_1 and S_2 could be substituted for H_1 and H_2 (S. Kim & Kolen, 2007; S. Kim & Lee, 2006).

Research Problem

Software programs have numerous options for specification of the prior information regarding the ability estimates and weights for these ability estimates. In addition, IRT programs, like Bilog-MG (Mislevy & Bock, 1990), produce estimates of the ability estimates and weights labeled as the quadrature points and posterior weights in the output which are commonly based on a standard normal distribution. Furthermore, a simulation by S. Kim and Lee (2006) used 100 equally spaced ability points from the range of -4 to 4 from the standard normal distribution. With many options returned by common software used for linking, it is imperative to understand the effects of the prior distributions and weights on the transformation constant estimates to guide practitioners.

To date, a single study by S. Kim and Kolen (2007) has explored the impact of distributional and symmetry related issues with the characteristic curve methods. This study found that mismatch between the ability distributions and the prior distributions resulted in more linking error compared to situations where the distributions matched. In addition, S. Kim and Kolen (2007) found that transformation constant estimates when using symmetric linking were between those using forward or backward linking (i.e. using only H_1 or H_2 respectively), suggesting that the symmetric linking is similar to taking an average of the criterion functions. Lastly, concurrent calibration was used as a comparison and was shown to perform better than the characteristic curve methods regardless of the distributional or symmetry options specified.

This study looks to extend more fully the S. Kim and Kolen (2007) study in at least three ways. First, the current study varies factors that were held constant in the S. Kim and Kolen (2007) study, including sample size, the percentage of common items, and assessment length. Secondly, the current study uses an R package, `plink` (R Core Team, 2015; Weeks, 2010), to conduct the linking in place of the POLYST software (S. Kim & Kolen, 2003). R is an open source statistical computing language and parallels between these study results and S. Kim and Kolen (2007) extends the generalizeability across two software implementations. Example code to perform the linking methods will also be provided. Finally, the primary outcome of interest is the bias of the transformation constants and the population transformation constants were varied as a part of the simulation design. This will allow for a greater extension of generalizeability of the simulation study conditions across a wide range of transformation constant values (Skrondal, 2000) and extend the work by S. Kim and Kolen (2007) in which their transformation constants were fixed with $A = 1$ and $B = 0$.

The research questions for this study include:

1. To what extent does the prior distribution have an impact on the estimation of the transformation constants?
2. To what extent does symmetric linking have an impact on the estimation of the transformation constants?
3. To what extent does the relationship from questions 1 & 2 generalize across the simulation conditions specified?

Methodology

A computer simulation was used to help answer the research questions. The following simulation conditions were manipulated, the ability distribution for form Y (4 levels: Normal(0, 1), Gamma(5, 2), $\chi^2(2)$, t(2)), the percentage of common items (3 levels: 10%, 25%, 40%), length of test (3 levels: 25, 40, 55 items), and sample size for Y (2 levels: 800, 1500). The choice of ability distributions was chosen to include both symmetric [Normal and t(2)] and skewed distributions [Gamma(5, 2) and $\chi^2(2)$] that may be present in real world assessment data. The percentage of common items and sample sizes were chosen to represent small sample conditions in which estimation problems would likely be exacerbated. In addition, the small percentage of common items may reflect situations in practice where items are tried out, but it is difficult to perform many tryout administrations. Therefore, the desired number of common items used for linking would be smaller and this study aims to explore those implications in more detail.

The simulation conditions discussed above were fully crossed in a factorial research design resulting in $4 * 3 * 3 * 2 = 72$ conditions. A total of 1000 replications for each of the 72 simulation conditions were performed. Note that the ability distributions for Form Y were centered based on the theoretical mean of each distribution to ensure they all have the same mean of zero. Differences in the standard deviations were retained to reflect differences in variability of the two groups. The variance of the ability distributions will be most extreme for the Gamma(5, 2) distribution with a theoretical variance of 20. Lastly, whenever the number of common items was a decimal (e.g. 2.5) it was always rounded up to the next whole number.

For each of the 72,000 generated data sets, two within conditions were fully crossed. These conditions represented the prior distribution of the old scale (3 levels) and new scale (4 levels). The three levels of the prior ability distribution for the old scale were 20 equally spaced proficiency points and weights between -4 to 4 from a uniform distribution, a standard normal distribution, and a Gamma(5, 2) distribution. Note that similar to the generated ability distributions above, the Gamma(5, 2) prior distribution was centered by subtracting the theoretical mean of 2.5. The four levels of the prior ability distribution for the new scale were no distribution (representing unsymmetric linking) and the same three distributions described above from the old scale. The Haebara and Stocking-Lord transformation methods were calculated for each within-study condition identified above.

The addition of the prior distribution resulted in another 12 within study condition factors for the characteristic curve methods. Note that the method moment linking methods will not vary for these within study conditions as the means and standard deviations of the IRT parameter estimates will not change. Therefore, there were a total of $72 * 24 = 1728$ characteristic study conditions and 144 method moment study conditions for a total of 1872 study conditions. A diagram highlighting the between and within study conditions can be seen in Figure 1.

In addition to these fully crossed simulation conditions, two additional conditions were simulated from a uniform distribution. These conditions were the transformation parameters that were the difference between the common item parameters for Form X and Form Y. More specifically, the A transformation constant was simulated from a random uniform distribution ranging from .5 to 1.5 rounded to the nearest .05, resulting in a total of 21 possibilities. Similarly, the B transformation constant was simulated from a random uniform distribution ranging from -2 to 2 rounded to the nearest tenth, resulting in a total of 41 possibilities. These two simulation conditions were not crossed with the above conditions, but instead were allowed to vary within each simulation condition. This was done to attempt to achieve stronger external validity of the simulation results across a variety of A and B population values.

Simulation Process

A population of 55 item parameters were simulated initially and represented a fixed set of items comprising Form X. A second form, Y, contained common items from Form X as well as new items where the item parameters were unknown and needed to be estimated. Based on the percentage of common items and the length of the test, the common items for Form Y were selected from the population of 55 fixed items from Form X. These were then transformed based on the A and B population transformation constants using the linking equations above [see Equation (1)]. Additional item parameters were randomly simulated to fill out the rest of Form Y. For example, if there were 3 common items on a 25 item assessment, the first 3 items were common with form X, the other 22 items were randomly simulated. Form Y will be linked onto the same scale as Form X. A diagram of the form simulation procedure can be seen in Figure 2.

After assembling the new form, individual responses were generated based on the item response functions and the individual's ability simulated from the four ability distributions described above. If the likelihood of answering the item was greater than a random uniform value between 0 and 1, the item was deemed to be answered correctly, otherwise answered incorrectly.

Analysis

After simulating the new responses for Form Y, these responses were calibrated with Bilog-MG (Mislevy & Bock, 1990). A 3 parameter IRT model was used for all calibrations. After calibration, transformation constants were calculated based on the items in common between the forms using an R package, `plink` (R Core Team, 2015; Weeks, 2010). These transformation constants were compared to the known population constant values to assess the recovery of these transformed values given the simulation conditions.

Bias of the transformation constants will be explored:

$$bias = \hat{\theta} - \theta \tag{10}$$

The bias formula above takes the difference from the parameter estimate ($\hat{\theta}$) and the parameter (θ) representing the raw difference between the parameter and the estimate. The expected value of this statistic if there is no bias would be 0. The bias will be explored for both the A and B transformation constants independently.

The bias will be explored descriptively and inferentially. Two linear models were used, one that contained all four linking methods (Haebara, Stocking-Lord, mean/mean, and mean/sigma) and a second model that restricted the analysis to only the characteristic curve methods. To identify simulation conditions that explained significant variation in the bias of the estimated transformation constants, effect sizes were used instead of hypothesis testing. The effect size used was η^2 and important simulation conditions were identified if the $\hat{\eta}^2$ effect size was greater than 0.001.

Variation in the bias statistics will also be explored using the root mean square error (RMSE) which takes the following form:

$$RMSE = \sqrt{\frac{\sum \left(\hat{\theta} - \theta\right)^2}{R}}, \tag{11}$$

where the numerator is the sum of the squared bias statistics within a study condition and R is the number of replications within the cell. The RMSE will be explored descriptively.

Verification of Simulation Values

Verification of the random uniform distribution for the A and B transformation constants was done to ensure adequate coverage across the range of values and uniformity across simulation conditions. Figure 3 depicts a heat map of the proportion of simulation conditions by the simulated B transformation constants. First, there does not appear to be any pattern to the plot for different rows of the plot, indicating that the simulated B transformation constants did not vary as a function of the simulation conditions.

Secondly, when looking at the different simulated B transformation constants (i.e. the columns), outside of the first and last columns (values of -2 and +2), there again does not appear to be differences across the range of transformation constants. The first and last columns represent areas where the rounding only had a range of 0.05 on the scale, for example from 1.95 to 2.00, instead of the rest having 0.10 of ranges to round to, for example 1.85 to 1.94 would round to 1.90. These smaller sample sizes were deemed to not be a concern due to the large number of replications, where if only .5% of the sample is found in a given cell, that would represent sample sizes of 360 and .1% of the sample represents 72 replications. A similar trend in the simulated A transformation constants was found (not shown).

Results

The bias was first explored descriptively. The average bias by the ability distribution for the method moment transformation methods is shown in Table 1. A similar table for the characteristic curve methods by the ability distribution and prior linking distribution is shown in Table 2. Both tables show descriptive differences in the average bias by the ability distributions and small differences between the methods. The bias of the A transformation constants also tend to be negative suggesting underestimation whereas the bias in the B transformation constants tend to be positive suggesting overestimation. Furthermore, the prior distributions for the old or new scale do not appear to have much impact on the average bias in the transformation constants shown in Table 2 within an ability distribution. There is some evidence that when using the Gamma(5, 2) prior distribution for the new scale that the average bias of the A transformation constants is slightly larger, but the prior distributions did not have a consistent pattern in the average bias for the B transformation constant within an ability distribution.

The standard deviations of these bias statistics are shown in parentheses within each table. For most methods, the variation in the bias of the B transformation constants is larger than the A transformation constants and is significantly larger for the skewed ability distribution [Chi-Sq (2)]. In addition, the t(2) ability distribution tended to increase the variation in the B transformation constants, but did not have as strong of an effect on the A transformation constants. Similar to the average bias, the Gamma(5, 2) prior distribution for the new scale increased the variation in the bias of the A transformation constants. The prior distributions did not appear to have an impact on the variation of the bias statistics for the B transformation constants within an ability distribution.

Lastly, correlations were used to see if the bias of the transformation constants were related to the simulated population transformation constant values. This analysis showed that there was no relationship between the population A and B values ($r = -0.001$), the A population constants and the bias of the B constants ($r = -0.01$), and the bias of the A and B constants ($r = 0.04$). Moderate to strong negative correlations were found between the A population constants and the bias of the A constants ($r = -0.36$) and between the B population constants and the bias of the B constants ($r = -0.70$). Finally, a small negative correlation was also found between the B population constants and the bias of the A constants ($r = -0.12$).

Root Mean Square Error

The RMSE was averaged across all 1000 replications and boxplots were created by combining RMSE values across the sample size, number of items, and number of common items simulation conditions. As a result, there were a total of 18 RMSE values in each boxplot within the charts. Figure 4 shows boxplots for the RMSE of the A transformation constant by the characteristic curve methods, the ability distribution, and prior distributions (both old and new scales). This figure reinforces the descriptive differences discussed above, particularly that the ability distribution accounts for much of the variation in the bias statistics for the A transformation constant. There are little to no differences across methods or across prior distributions for the new or old scale. Also, the variation across the 18 RMSE value tends to be small for each condition shown in Figure 4, with no discernible pattern in those that have larger amounts of variation.

Figure 5 shows RMSE boxplots for the B transformation constants by the same predictors as the previous figure. This figure shows much smaller gaps between the four ability distributions and the RMSE values tend to cluster into two categories. One category tends to represent the Chi-Square (2) and Gamma(5, 2) distributions and the second represent the t(2) and normal distributions. This is not surprising as the first two are non-symmetric distributions whereas the second two are symmetric distributions. The symmetric ability distributions tend to have much smaller RMSE values compared to the unsymmetric ability distributions. The unsymmetric linking method with a Gamma(5, 2) prior distribution for the old scale has much larger variation than the other conditions. The other conditions tended to have small amounts of variation in the RMSE and were consistent across the conditions.

Inferential Statistics

As there was variation in the bias of the transformation constants, the simulation conditions were used to attempt to explain variation in the bias of the transformation constants. The linking method, ability distribution, prior linking distribution (old and new), sample size, number of items, and number of common items were treated as factors and the A and B population constants were entered as continuous variables in the model. The A population values were centered at zero (i.e. one was subtracted from all A population values) to ensure the intercept is interpreted as values when the IRT parameters for the common items are already on the same scale (i.e. the A and B transformation constants are both equal to 0).

Effect sizes larger than 0.001 for all linking methods can be seen in Table 3. Overall, the models with three-way interactions explained 83% and 93% of the variation in the bias of the A and B transformation estimates respectively; while the effect sizes shown in Table 3 explained about 81% and 92% of the variation respectively. Sample size, number of items, and number of common items did not help explain variation in the bias for either transformation constant which suggests that bias is not affected by these factors. The variables that explained the most variation in the bias of the A and B transformation constants were the ability distribution, transformation constant population values, and the interaction between the two. Only a small amount of variation in the bias of the transformation constants was explained by the old and new prior distributions, less than 1% in all instances.

The linking method explained small amounts of variation in the bias of the transformation constants, however the effect was much smaller compared to the other effects discussed. The average effects and standard deviations for the four methods can be seen in Table 5 where small differences between the methods can be seen, particularly, the mean/mean method had slightly larger average bias for the A and B transformation estimates. In a few instances, the mean/sigma method had bias statistics closer to zero, however there was evidence that these were much more variable, especially when the ability distribution was normally distributed. The larger effect is shown in the rows of the table, with large differences in the average bias statistics across ability distributions. In general, the Chi-Square (2) distributions had the largest average bias statistics for both transformation constants across methods. The least amount of bias was found when the ability distribution was normally distributed and the Gamma(5, 2) distribution performed better than the t(2) distribution.

Characteristic Curve Methods Only

A second set of models were run that omitted the method moment linking methods and only analyzed the characteristic curve methods. The effect sizes larger than 0.001 can be seen in Table 4 and these explained approximately 84% and 92% of the variation in the bias of the A and B transformation constants respectively. The method variable no longer explained variation when the analysis was restricted to the characteristic curve methods, suggesting no meaningful difference between the estimates from Haebara and Stocking-Lord methods. The prior distributions, old and new scale, explained small amounts of variations, less than 1% for both the A and B transformation constants.

The average effects by the ability distribution (different lines and shapes), prior distribution for the old scale (facets of the plot), and the A and B population values can be seen in Figure 6 and Figure 7 respectively. As can be seen from the figures, the least amount of bias occurs when the ability distribution is distributed as either normal or Gamma(5, 2). When the ability distribution is skewed [Chi-Square (2)] or has heavy tails [t(2)] the amount of bias increases significantly. In addition, the prior distribution for the old scale has a very small effect on the bias, however, in most conditions, the Gamma(5, 2) distribution performs the worst, even when this matches the ability distribution. This suggests that on average using either a normal or uniform prior distribution is sufficient.

Finally, the A transformation estimates tend to be underestimated when bias occurs, whereas with the B transformation estimates, bias is more strongly a function of the actual B population values. For example, the B transformation estimates tend to be overestimated when the B population values are negative and underestimated when the B population values are positive. This helps to visualize the negative correlation found in the descriptive analysis between these two variables and provides evidence this persists even when other simulation factors are accounted for. Finally, considering the scale for the bias statistics also suggests that bias may be an extreme concern in some situations. For example, when the ability distribution is Chi-Square (2) and the A population value is 1.5, the bias approaches -1.0, suggesting that the estimated A transformation constant was close to 0.5. A similar finding is found for the same conditions with the B transformation constant when the B population value is close to 2 in absolute value, the average bias is greater than 1 in absolute value. This suggests that the transformation constants are significantly underestimated for the Chi-Square (2) ability distribution and could have strong impacts on the item parameters used in an assessment.

Empirical Example

An empirical data example is used to show the differences for the different linking methods while specifying different proficiency points and weights. Data from non-equivalent groups came from a 1992 administration of the Iowa Tests of Basic Skills vocabulary assessment from grades 3 through 8. These data were used in an article by Tong and Kolen (2007) and can be found in the R package plink (Weeks, 2010). The code to do these linking computations with the plink R package can be seen in Appendix A.

Twelve different combinations of weights were used on the empirical data to show the differences in estimates of the characteristic curve linking methods. Three combinations included normal, uniform, and Gamma(5, 2) proficiency points and weights for the old scale without considering the new scale representing non-symmetric linking designs. To explore differences when symmetry is assumed, the same proficiency points and weights, [normal, uniform, and Gamma(5, 2)], were used for the new scale. All combinations were explored resulting in an additional nine combinations (see Appendix A for the syntax for these twelve linking designs). Grade three was used as the base group (i.e. IRT parameters were placed on the grade three scale).

Table 6 shows the transformation constants to create a vertical scale for the item parameters by way of linking. The rows of the table represent different specification of the weights for the old and new metric. As can be seen from the table, the differences are modest within each linking method and weights old, however there are small differences associated with the method. Particularly, the Haebara linking constants tend to be slightly smaller in many conditions compared to the Stocking-Lord constants. These differences are very modest and are likely not of practical importance to practitioners. In addition, the Gamma(5, 2) distribution for the proficiency points and weights for the old scale, produces smaller estimates going from grade 7 to grade 8 (the last column of Table 6).

Discussion

This study explored the effect of using different prior distributions (i.e. proficiency points and weights) when linking the old and new ability scale with characteristic curve linking methods. These methods are common when two (or more) assessments have been given to non-equivalent groups with common or anchor items. Through the use of common items, differences in the item characteristic curves or test characteristic curves are minimized through a linear transformation [see Equation (1)]. However, to date, only a single study by S. Kim and Kolen (2007) have explored the implications for different specifications of these prior distributions of the old and new scale. The current study aimed to expand on this literature by using another software implementation of the linking routines and also by varying additional simulation conditions.

The results showed that the prior distribution for the old and new scales are not an important consideration when looking at bias of the transformation constants and generalizes across the simulation conditions considered. For example, the effect due to prior distribution was similar across the number of common items, sample size, linking method, or even the transformation population values. This provides some reassurances to researchers or practitioners that the results are not sensitive to the prior distribution and even provides evidence that using simple prior distributions such as a normal or uniform distribution provide similar results. For example, a study by S. Kim and Lee (2006) used a uniform prior distribution when using characteristic curve linking methods. These results do differ slightly from the S. Kim and Kolen (2007) study where they found poorer performance when the ability and prior distributions did not match. However, a similar finding was found in that ability distributions that were normally distributed provided the best performance.

The prior distribution also did not interact with the linking method, suggesting that the choice of the characteristic curve methods are robust with regard to the prior distributions. Similar to prior work, method moment methods showed larger amounts of bias compared to the characteristic curve methods (Hanson & Béguin, 2002; S. Kim & Lee, 2006), suggesting that the method moment methods are less desirable than the characteristic curve methods. With the improvements in modern computers, the computations needed to perform the characteristic curve linking methods are very quick to perform.

The larger problem found in this simulation study was with respect to skewed or heavy tailed ability distributions which can be common in real world data (Micceri, 1989). In these situations, the transformation constants can have significant amounts of bias with larger variation in their estimates, a similar finding found by S. Kim and Kolen (2007). Perhaps additional prior distributions can help to alleviate these concerns or a normalizing transformation of the ability distribution before performing the linking may be helpful. Further research into these areas would be useful to better recommend alternatives for those using these methods with non-normal ability distributions.

Lastly, the number of common items did not significantly impact the estimates of the A and B transformation constants. The smallest number of common items considered in this study were three items, which suggests that on average there is no difference in the average bias when using three common items or ten. Increased error in the linking methods has been shown by others when using a small number of common items to conduct the linking (Hanson & Béguin, 2002), which was shown by the increased variation in the transformation constant estimates with few common items. The mean/sigma method was one instance that had increased variation in the bias with only three common items compared to larger number of common items. This is not surprising given the estimate of the standard deviation of the item difficulties would be much more variable.

Conclusions

In practice, these results suggest that the prior distribution used for the linking of two forms is not an aspect that researchers or practitioners need to concern themselves with. Of more concern, is understanding the shape of the ability distribution of the respondents. If there is evidence or concern that respondents do not follow an approximately normal ability distribution, extreme care needs to be taken when linking these two forms. In these cases, bias in the transformation constants can provide a poor link between the two forms. The ability distribution does not simply need to be symmetric, care in exploring the tails of the distribution is also needed as this study showed that ability distributions that are symmetric with heavier tails [i.e. t(2)] pose estimation concerns for the transformation constants.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores.*

Camilli, G. (1994). Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, *19*(3), 293–295.

De Ayala, R. J. (2013). *Theory and practice of item response theory.* Guilford Publications.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating test scores. *Educational measurement*, *4*, 187–220.

Kim, J. & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, *26*(3), 255–270.

Kim, S. & Kolen, M. J. (2003). *POLYST: A computer program for polytomous IRT scale transformation.* Iowa City: Iowa Testing Programs, The University of Iowa. Retrieved from http://www.education.uiowa.edu/casma

Kim, S. & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the irt characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371–397.

Kim, S. & Lee, W.-C. (2006). An extension of four irt linking methods for mixed-format tests. *Journal of Educational Measurement*, *43*(1), 53–76.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking.* Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*(1), 156.

Mislevy, R. J. & Bock, R. D. (1990). *Bilog 3: item analysis and test scoring with binary logistic models*. Scientific Software International.

R Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Skrondal, A. (2000). Design and analysis of monte carlo experiments: attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201–210.

Tong, Y. & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*(2), 227–253.

Weeks, J. P. (2010). plink: an R package for linking mixed-format tests using irt-based methods. *Journal of Statistical Software*, *35*(12), 1–33. Retrieved from http://www.jstatsoft.org/v35/i12/

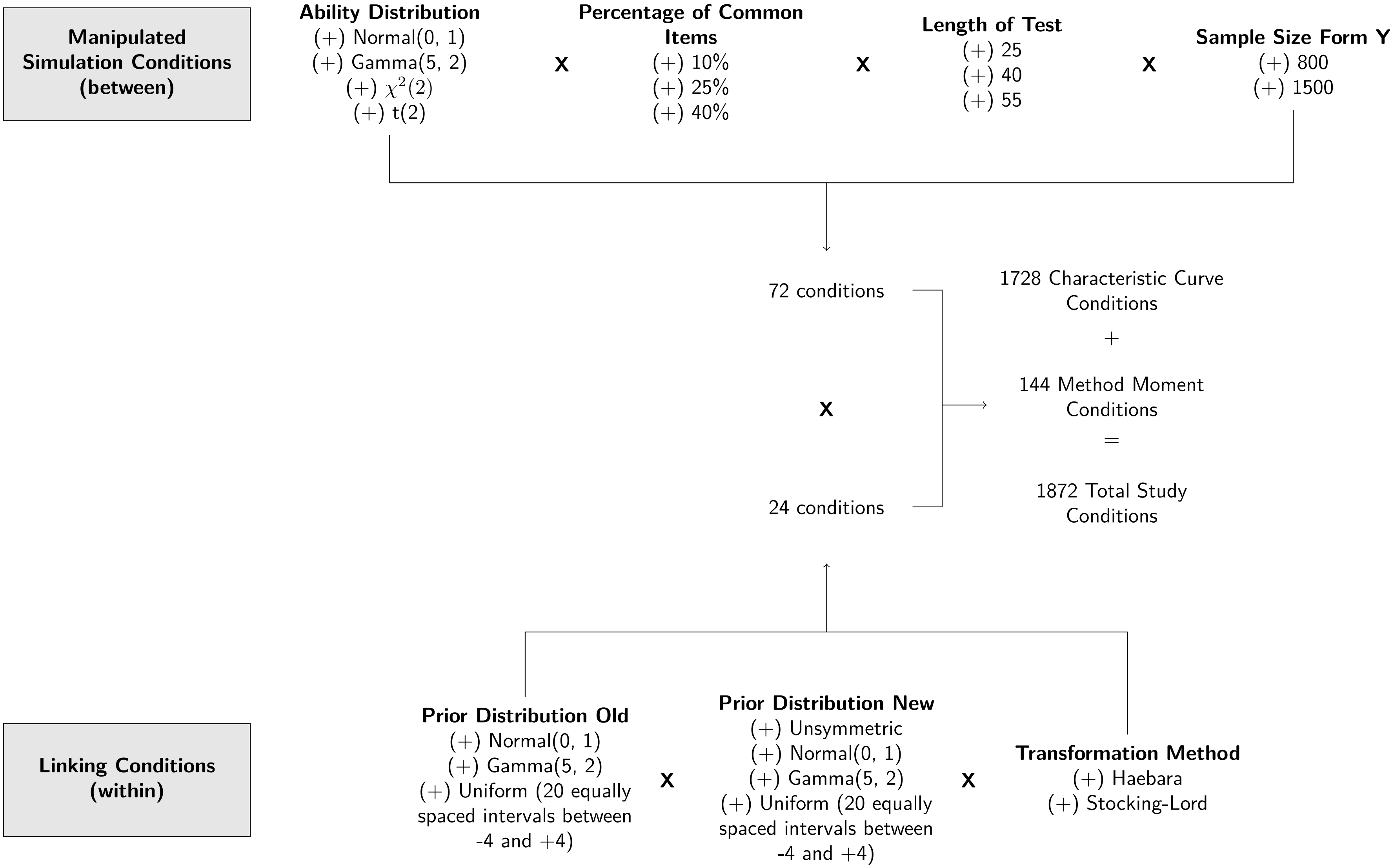

Figure 1: Diagram depicting simulation conditions considered without varied A and B transformation constants.

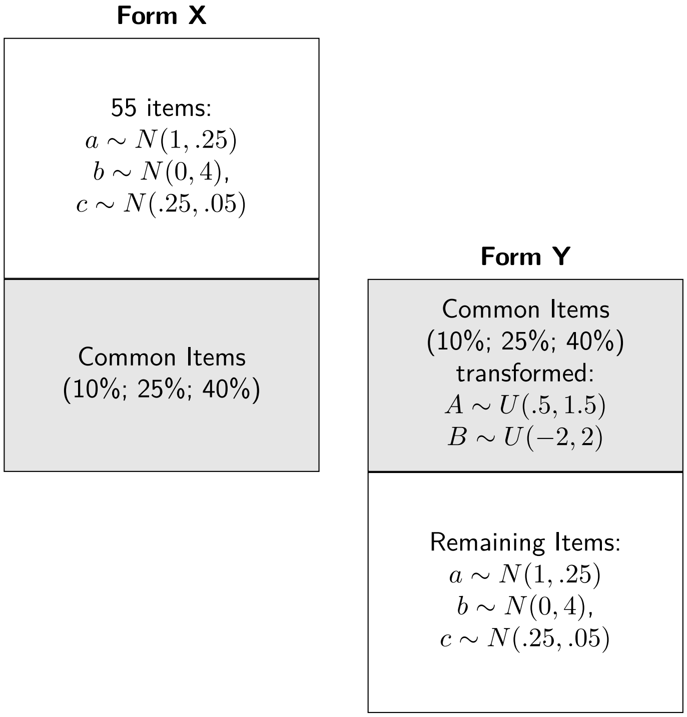

Figure 2: Diagram depicting simulation of item responses for the two forms.

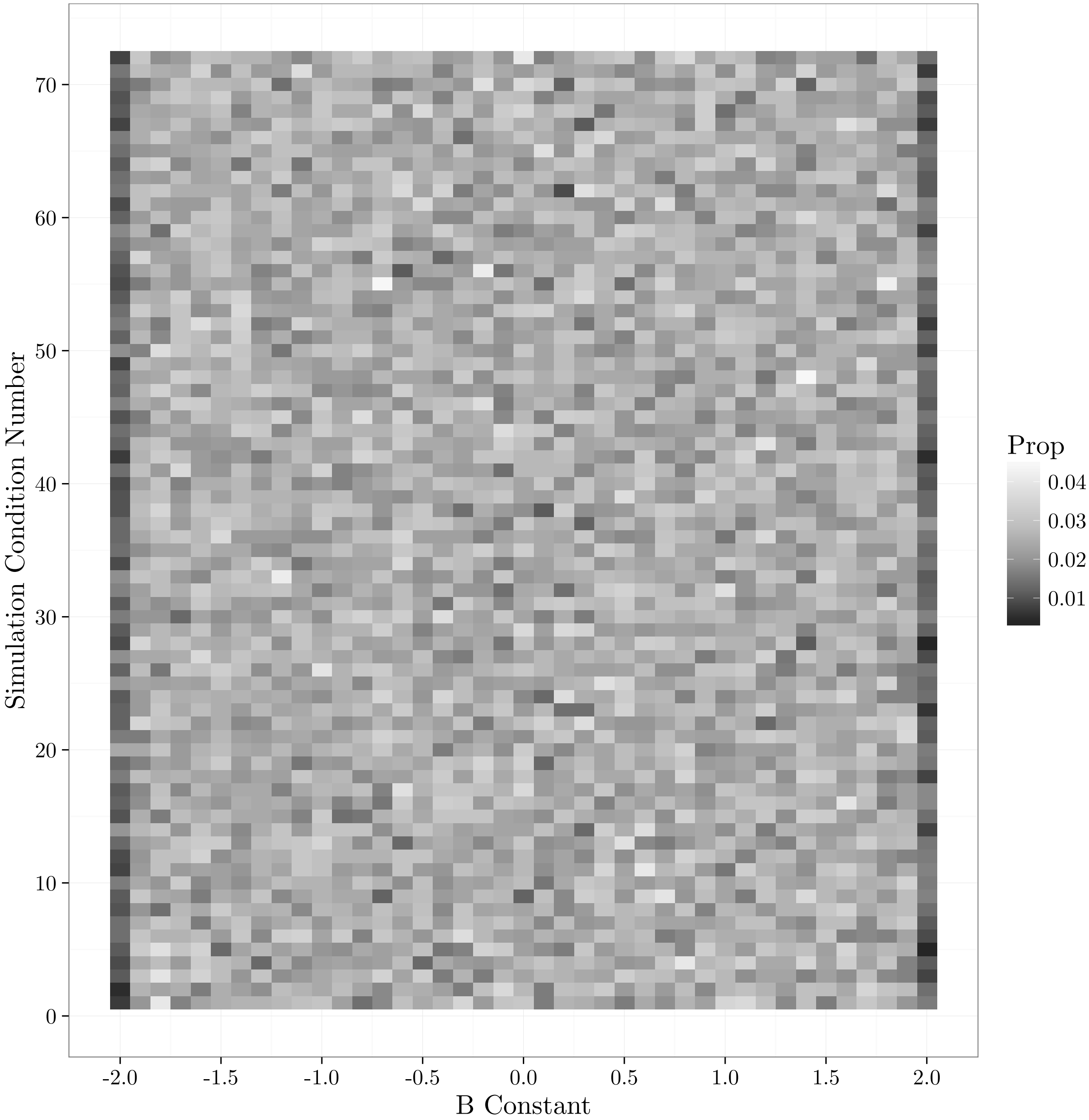

Figure 3: Heatmap showing simulation condition coverage across the range of simulation values for the B population transformation constants.

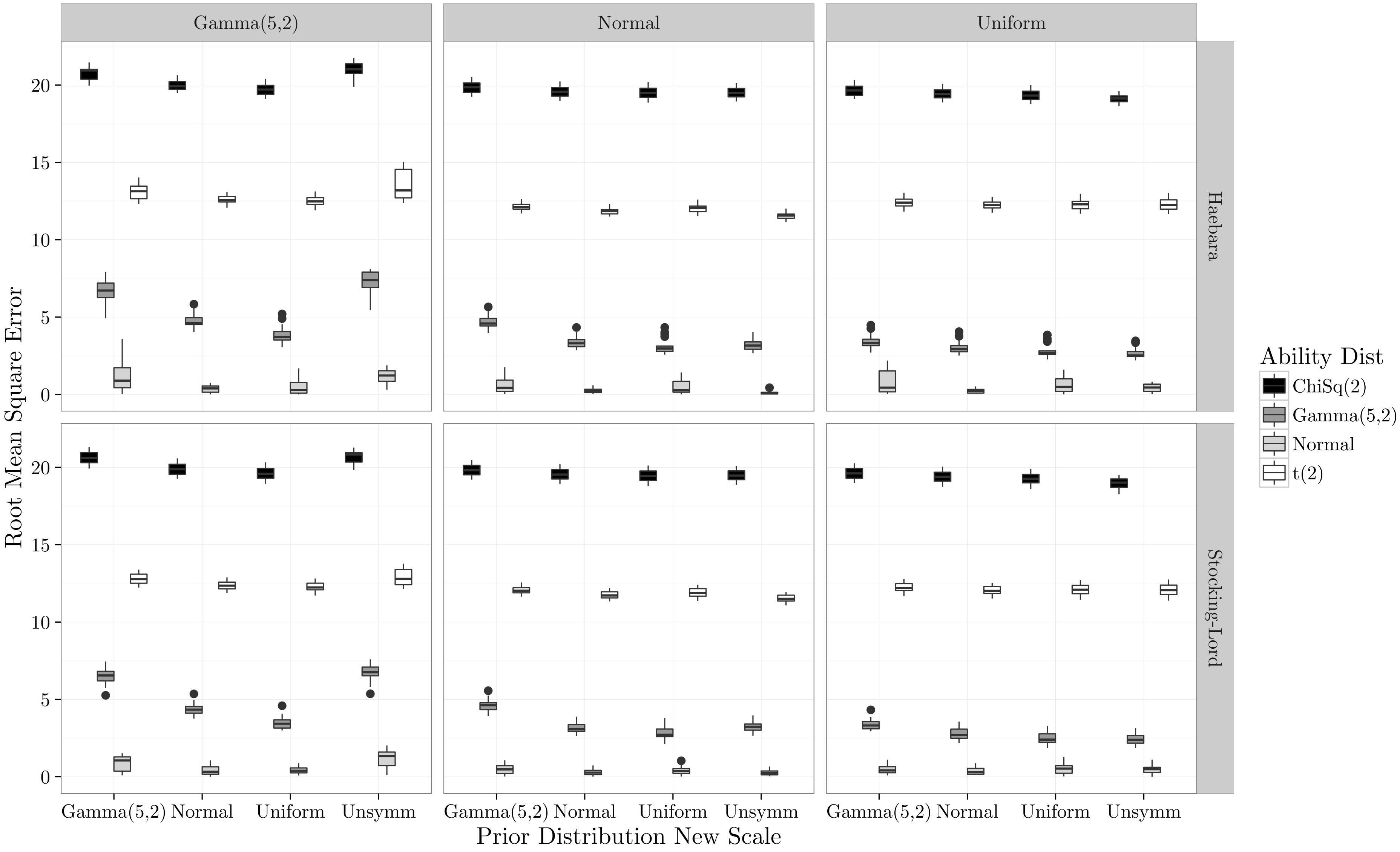

Figure 4: Boxplots of the root mean square error of the A transformation constant by the prior distribution of the new scale (x-axis) and old scale (vertical facet), method (horizontal panels), and ability distribution.

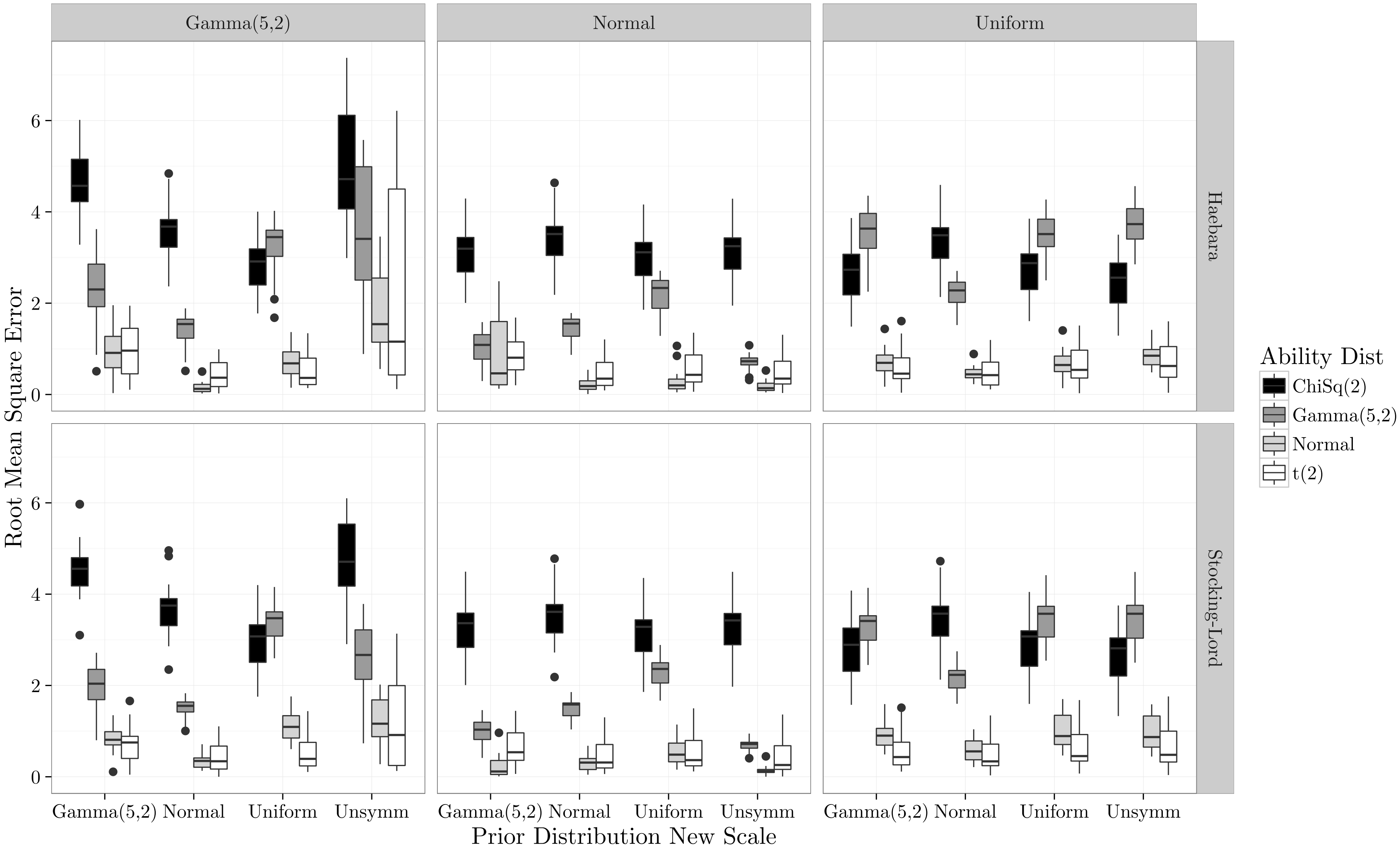

Figure 5: Boxplots of the root mean square error of the B transformation constant by the prior distribution of the new scale (x-axis) and old scale (vertical facet), method (horizontal panels), and ability distribution.

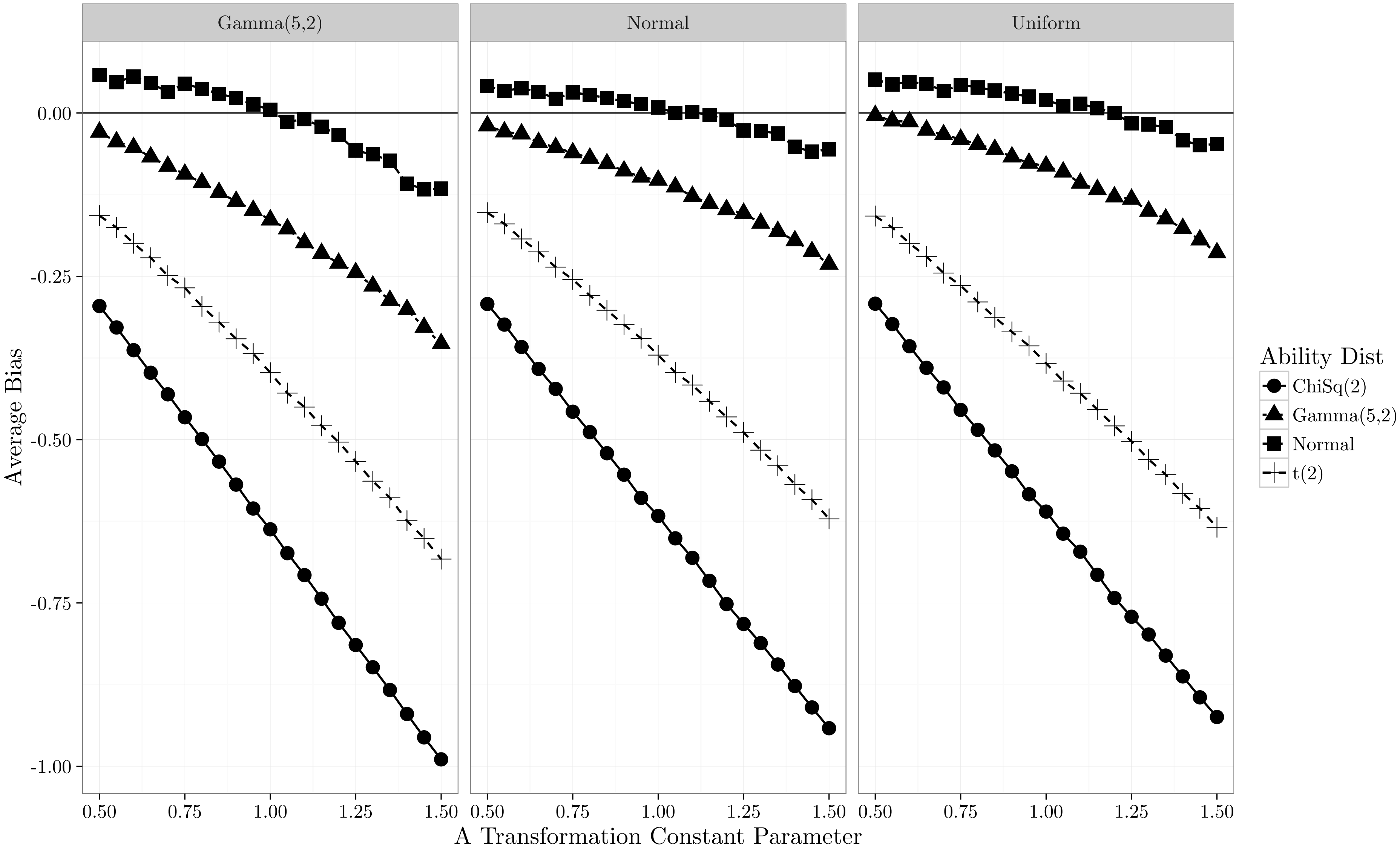

Figure 6: Bias of the A transformation constant by A population values, ability distribution, and linking prior distribution for the old scale.

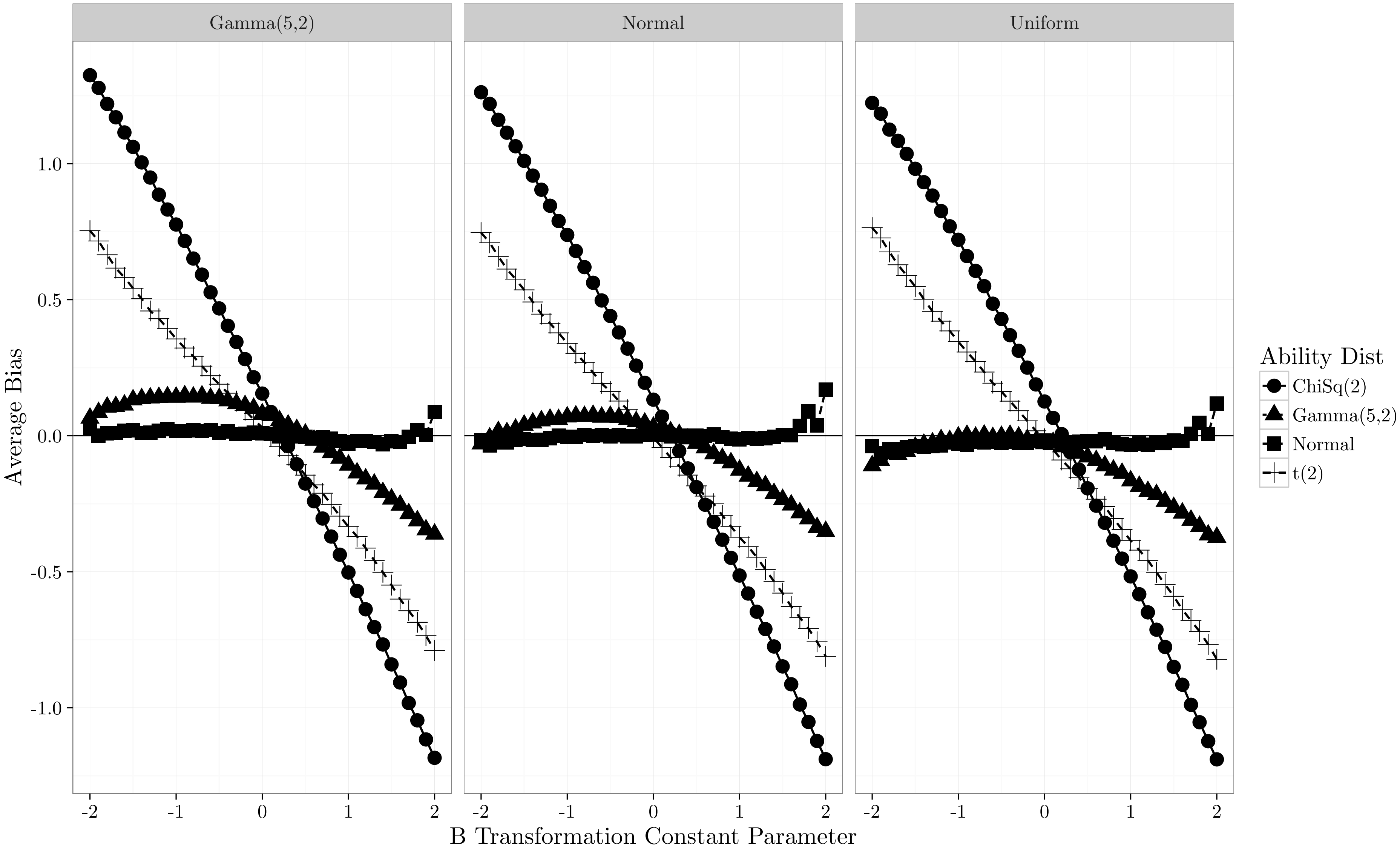

Figure 7: Bias of the B transformation constant by B population values, ability distribution, and linking prior distribution for the old scale.

Table 1: Mean (SDs) for bias of transformation constants for method moment transformation methods

	Mean/Mean		Mean/Sigma	
Ability Distribution	A	B	A	B
ChiSq(2)	-0.644 (0.21)	0.148 (0.74)	-0.629 (0.21)	0.148 (0.74)
Gamma(5,2)	-0.202 (0.15)	0.076 (0.20)	-0.084 (0.25)	0.058 (0.18)
Normal	-0.055 (0.13)	0.093 (0.35)	0.086 (0.61)	0.068 (0.23)
t(2)	-0.398 (0.17)	-0.013 (0.42)	-0.361 (0.16)	-0.019 (0.42)

Table 2: Mean (SDs) for bias of transformation constants for characteristic curve transformation methods by the ability distribution, prior distribution for the old and new scale.

			Haebara		Stocking-Lord	
Ability Dist	Prior Dist Old	Prior Dist New	A	B	A	B
ChiSq(2)	Normal	Gamma(5,2)	-0.627 (0.20)	0.099 (0.72)	-0.626 (0.20)	0.103 (0.72)
ChiSq(2)	Normal	Unsymm	-0.617 (0.19)	0.099 (0.72)	-0.616 (0.19)	0.104 (0.72)
ChiSq(2)	Normal	Normal	-0.619 (0.19)	0.108 (0.72)	-0.618 (0.19)	0.111 (0.72)
ChiSq(2)	Normal	Uniform	-0.617 (0.19)	0.095 (0.71)	-0.615 (0.19)	0.100 (0.71)
ChiSq(2)	Uniform	Gamma(5,2)	-0.622 (0.20)	0.085 (0.71)	-0.620 (0.19)	0.089 (0.71)
ChiSq(2)	Uniform	Unsymm	-0.605 (0.19)	0.078 (0.70)	-0.601 (0.19)	0.084 (0.70)
ChiSq(2)	Uniform	Normal	-0.615 (0.19)	0.107 (0.72)	-0.613 (0.19)	0.109 (0.72)
ChiSq(2)	Uniform	Uniform	-0.612 (0.19)	0.087 (0.71)	-0.609 (0.19)	0.091 (0.71)
ChiSq(2)	Gamma(5,2)	Gamma(5,2)	-0.657 (0.21)	0.149 (0.75)	-0.653 (0.21)	0.145 (0.74)
ChiSq(2)	Gamma(5,2)	Unsymm	-0.663 (0.21)	0.162 (0.76)	-0.654 (0.21)	0.151 (0.75)
ChiSq(2)	Gamma(5,2)	Normal	-0.631 (0.20)	0.113 (0.72)	-0.628 (0.20)	0.115 (0.72)
ChiSq(2)	Gamma(5,2)	Uniform	-0.623 (0.20)	0.090 (0.71)	-0.620 (0.20)	0.094 (0.71)
Gamma(5,2)	Normal	Gamma(5,2)	-0.148 (0.12)	-0.033 (0.17)	-0.146 (0.12)	-0.032 (0.17)
Gamma(5,2)	Normal	Unsymm	-0.101 (0.09)	-0.022 (0.14)	-0.102 (0.10)	-0.022 (0.14)
Gamma(5,2)	Normal	Normal	-0.107 (0.13)	-0.046 (0.15)	-0.099 (0.13)	-0.047 (0.15)
Gamma(5,2)	Normal	Uniform	-0.098 (0.13)	-0.070 (0.15)	-0.089 (0.14)	-0.073 (0.14)
Gamma(5,2)	Uniform	Gamma(5,2)	-0.109 (0.13)	-0.112 (0.15)	-0.107 (0.13)	-0.105 (0.15)
Gamma(5,2)	Uniform	Unsymm	-0.085 (0.13)	-0.118 (0.14)	-0.077 (0.13)	-0.110 (0.14)
Gamma(5,2)	Uniform	Normal	-0.097 (0.14)	-0.070 (0.15)	-0.087 (0.14)	-0.069 (0.14)
Gamma(5,2)	Uniform	Uniform	-0.089 (0.13)	-0.110 (0.14)	-0.078 (0.14)	-0.110 (0.14)
Gamma(5,2)	Gamma(5,2)	Gamma(5,2)	-0.212 (0.15)	0.073 (0.25)	-0.206 (0.14)	0.062 (0.22)
Gamma(5,2)	Gamma(5,2)	Unsymm	-0.230 (0.15)	0.112 (0.29)	-0.214 (0.14)	0.082 (0.24)
Gamma(5,2)	Gamma(5,2)	Normal	-0.149 (0.15)	-0.044 (0.15)	-0.138 (0.15)	-0.048 (0.15)
Gamma(5,2)	Gamma(5,2)	Uniform	-0.122 (0.14)	-0.101 (0.14)	-0.110 (0.15)	-0.107 (0.14)
Normal	Normal	Gamma(5,2)	0.010 (0.38)	0.029 (0.53)	-0.012 (0.20)	0.005 (0.22)
Normal	Normal	Unsymm	0.002 (0.08)	0.005 (0.11)	-0.001 (0.10)	0.003 (0.12)
Normal	Normal	Normal	-0.007 (0.09)	-0.007 (0.11)	-0.002 (0.11)	-0.010 (0.12)
Normal	Normal	Uniform	0.015 (0.22)	-0.003 (0.24)	0.009 (0.14)	-0.017 (0.15)
Normal	Uniform	Gamma(5,2)	0.026 (0.36)	-0.014 (0.41)	0.006 (0.19)	-0.029 (0.18)
Normal	Uniform	Unsymm	0.013 (0.10)	-0.027 (0.13)	0.011 (0.11)	-0.032 (0.12)
Normal	Uniform	Normal	0.004 (0.10)	-0.015 (0.12)	0.006 (0.11)	-0.019 (0.12)
Normal	Uniform	Uniform	0.019 (0.20)	-0.021 (0.20)	0.014 (0.14)	-0.033 (0.14)
Normal	Gamma(5,2)	Gamma(5,2)	0.017 (0.58)	0.009 (0.26)	-0.023 (0.26)	0.019 (0.15)
Normal	Gamma(5,2)	Unsymm	-0.029 (0.16)	0.042 (0.19)	-0.035 (0.14)	0.028 (0.15)
Normal	Gamma(5,2)	Normal	-0.011 (0.12)	-0.004 (0.12)	-0.013 (0.12)	-0.011 (0.12)
Normal	Gamma(5,2)	Uniform	0.014 (0.31)	-0.022 (0.16)	0.005 (0.15)	-0.035 (0.13)
t(2)	Normal	Gamma(5,2)	-0.383 (0.15)	-0.025 (0.44)	-0.381 (0.15)	-0.020 (0.44)
t(2)	Normal	Unsymm	-0.365 (0.14)	-0.012 (0.43)	-0.364 (0.14)	-0.010 (0.42)
t(2)	Normal	Normal	-0.374 (0.14)	-0.011 (0.43)	-0.372 (0.14)	-0.009 (0.43)
t(2)	Normal	Uniform	-0.381 (0.15)	-0.015 (0.45)	-0.376 (0.14)	-0.014 (0.44)
t(2)	Uniform	Gamma(5,2)	-0.393 (0.16)	-0.016 (0.45)	-0.387 (0.15)	-0.016 (0.44)
t(2)	Uniform	Unsymm	-0.388 (0.15)	-0.022 (0.46)	-0.381 (0.14)	-0.020 (0.44)
t(2)	Uniform	Normal	-0.387 (0.15)	-0.012 (0.43)	-0.381 (0.15)	-0.011 (0.43)
t(2)	Uniform	Uniform	-0.388 (0.15)	-0.019 (0.46)	-0.381 (0.15)	-0.019 (0.44)
t(2)	Gamma(5,2)	Gamma(5,2)	-0.415 (0.19)	0.026 (0.41)	-0.406 (0.18)	0.017 (0.41)
t(2)	Gamma(5,2)	Unsymm	-0.428 (0.18)	0.064 (0.47)	-0.408 (0.18)	0.028 (0.43)
t(2)	Gamma(5,2)	Normal	-0.399 (0.16)	-0.007 (0.43)	-0.391 (0.16)	-0.007 (0.42)
t(2)	Gamma(5,2)	Uniform	-0.396 (0.16)	-0.013 (0.45)	-0.388 (0.15)	-0.014 (0.44)

Table 3: Effect sizes from analysis of variance model exploring absolute bias of A and B parameters for characteristic curve and method moment linking methods.

Variable	$\hat{\eta}^2$ A	$\hat{\eta}^2$ B
Method	0.003	0.003
Ability Dist.	0.583	0.015
Prior Dist. Old	0.002	0.002
A Pop.	0.124	
B Pop.	0.014	0.474
Method:Ability Dist.	0.002	0.002
Ability Dist.:Prior Dist. Old	0.001	
Ability Dist.:A Pop.	0.036	
Ability Dist.:B Pop.	0.012	0.404
A Pop.:B Pop.	0.002	0.003
Method:Ability Dist.:B Pop.	0.002	
Method:A Pop.		0.001
Prior Dist. Old:Prior Dist. New		0.001
Ability Dist.:A Pop.:B Pop.		0.001

Note: “:” = interaction, Old = Old Scale, New = New Scale

Table 4: Effect sizes from analysis of variance model exploring absolute bias of A and B parameters for only characteristic curve methods.

Variable	$\hat{\eta}^2$ A	$\hat{\eta}^2$ B
Ability Dist.	0.640	0.017
Prior Dist. Old	0.003	0.002
A Pop.	0.132	
B Pop.	0.015	0.488
Ability Dist.:Prior Dist. Old	0.001	
Ability Dist.:A Pop.	0.040	
Ability Dist.:B Pop.	0.011	0.404
Prior Dist. Old:A Pop.	0.001	
A Pop.:B Pop.	0.001	0.003
Prior Dist. Old:Prior Dist. New		0.002
Ability Dist.:A Pop.:B Pop.		0.001

Note: ":" = interaction, Old = Old Scale, New = New Scale

Table 5: Means (SDs) for significant method by ability distribution interaction for bias of transformation constants.

	Haebara		Mean/Mean		Mean/Sigma		Stocking Lord	
Ability Dist.	A	B	A	B	A	B	A	B
ChiSq(2)	-0.628 (0.20)	0.113 (0.73)	-0.644 (0.21)	0.148 (0.74)	-0.629 (0.21)	0.148 (0.74)	-0.624 (0.20)	0.113 (0.73)
Gamma(5,2)	-0.139 (0.14)	-0.009 (0.22)	-0.202 (0.15)	0.076 (0.20)	-0.084 (0.25)	0.058 (0.18)	-0.131 (0.14)	-0.017 (0.20)
Normal	-0.005 (0.12)	0.007 (0.15)	-0.055 (0.13)	0.093 (0.35)	0.086 (0.61)	0.068 (0.23)	-0.008 (0.12)	-0.000 (0.13)
t(2)	-0.394 (0.16)	0.010 (0.45)	-0.398 (0.17)	-0.013 (0.42)	-0.361 (0.16)	-0.019 (0.42)	-0.384 (0.15)	-0.000 (0.43)

Table 6: Linking constant values by linking method and prior distribution for old and new scale by grade.

Method	Prior Dist Old	Prior Dist New	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
SL	Uniform	Gamma(5, 2)	1.030, 0.522	1.090, 0.381	1.069, 0.294	1.070, 0.165	1.079, 0.156
SL	Uniform	Normal	1.035, 0.527	1.093, 0.383	1.068, 0.299	1.073, 0.174	1.072, 0.150
SL	Uniform	Uniform	1.033, 0.524	1.091, 0.382	1.074, 0.296	1.079, 0.171	1.065, 0.156
SL	Uniform	None	1.032, 0.524	1.091, 0.382	1.074, 0.296	1.079, 0.171	1.066, 0.156
SL	Normal	Gamma(5, 2)	1.035, 0.525	1.093, 0.382	1.062, 0.299	1.064, 0.173	1.084, 0.148
SL	Normal	Normal	1.040, 0.529	1.096, 0.384	1.063, 0.300	1.068, 0.175	1.075, 0.147
SL	Normal	Uniform	1.038, 0.528	1.094, 0.383	1.067, 0.299	1.074, 0.175	1.070, 0.150
SL	Normal	None	1.043, 0.530	1.097, 0.384	1.064, 0.300	1.070, 0.176	1.074, 0.147
SL	Gamma(5, 2)	Gamma(5, 2)	1.020, 0.533	1.084, 0.390	1.059, 0.302	1.055, 0.169	1.119, 0.117
SL	Gamma(5, 2)	Normal	1.020, 0.533	1.084, 0.390	1.059, 0.302	1.055, 0.169	1.119, 0.117
SL	Gamma(5, 2)	Uniform	1.020, 0.533	1.084, 0.390	1.059, 0.302	1.055, 0.169	1.119, 0.117
SL	Gamma(5, 2)	None	1.019, 0.534	1.083, 0.390	1.060, 0.301	1.052, 0.173	1.117, 0.119
HB	Uniform	Gamma(5, 2)	1.026, 0.507	1.089, 0.377	1.069, 0.301	1.082, 0.159	1.100, 0.148
HB	Uniform	Normal	1.031, 0.519	1.092, 0.378	1.066, 0.302	1.078, 0.170	1.087, 0.145
HB	Uniform	Uniform	1.027, 0.514	1.085, 0.377	1.068, 0.301	1.090, 0.168	1.083, 0.150
HB	Uniform	None	1.032, 0.513	1.088, 0.376	1.069, 0.300	1.089, 0.167	1.086, 0.148
HB	Normal	Gamma(5, 2)	1.027, 0.516	1.093, 0.378	1.065, 0.303	1.070, 0.170	1.101, 0.143
HB	Normal	Normal	1.034, 0.522	1.097, 0.380	1.063, 0.302	1.071, 0.172	1.087, 0.143
HB	Normal	Uniform	1.030, 0.520	1.090, 0.379	1.064, 0.302	1.080, 0.172	1.084, 0.145
HB	Normal	None	1.039, 0.525	1.099, 0.381	1.061, 0.302	1.073, 0.172	1.087, 0.142
HB	Gamma(5, 2)	Gamma(5, 2)	1.009, 0.519	1.078, 0.388	1.065, 0.304	1.098, 0.117	1.143, 0.098
HB	Gamma(5, 2)	Normal	1.009, 0.519	1.078, 0.388	1.065, 0.304	1.098, 0.117	1.143, 0.098
HB	Gamma(5, 2)	Uniform	1.009, 0.519	1.078, 0.388	1.065, 0.304	1.098, 0.117	1.143, 0.098
HB	Gamma(5, 2)	None	1.002, 0.525	1.070, 0.395	1.063, 0.305	1.079, 0.140	1.131, 0.109

Note: SL = Stocking-Lord; HB = Haebara; Constant values specified as A, B; Base Group is Grade 3

Appendix A

```
library(plink)


pars <- TK07$pars
common <- TK07$common
cat <- list(rep(2,26),rep(2,34),rep(2,37),rep(2,40),rep(2,41),rep(2,43))
pm1 <- as.poly.mod(26)
pm2 <- as.poly.mod(34)
pm3 <- as.poly.mod(37)
pm4 <- as.poly.mod(40)
pm5 <- as.poly.mod(41)
pm6 <- as.poly.mod(43)
pm <- list(pm1, pm2, pm3, pm4, pm5, pm6)
x <- as.irt.pars(pars, common, cat, pm,
                 grp.names=paste("grade",3:8,sep=""))

# Old Scale weights only
unif_to <- plink(x, weights.t = as.weight(20))
norm_to <- plink(x, weights.t = as.weight(20, quadrature = TRUE))
gamma_to <- plink(x, weights.t = as.weight(20, quadrature = TRUE,
                  dist = 'gamma', alpha = 5, beta = 1/2))

# Old Scale weights with unif New Scale weights
unif_unif <- plink(x, weights.t = as.weight(20),
                   weights.f = as.weight(20), symmetric = TRUE)
norm_unif <- plink(x, weights.t = as.weight(20, quadrature = TRUE),
                   weights.f = as.weight(20), symmetric = TRUE)
gamma_unif <- plink(x, weights.t = as.weight(20, quadrature = TRUE,
                    dist = 'gamma', alpha = 5, beta = 1/2),
                    weight.f = as.weight(20), symmetric = TRUE)

# Old Scale weights with normal New Scale weights
```

```
unif_norm <- plink(x, weights.t = as.weight(20),
                   weights.f = as.weight(20, quadrature = TRUE), symmetric = TRUE)
norm_norm <- plink(x, weights.t = as.weight(20, quadrature = TRUE),
                   weights.f = as.weight(20, quadrature = TRUE), symmetric = TRUE)
gamma_norm <- plink(x, weights.t = as.weight(20, quadrature = TRUE,
                    dist = 'gamma', alpha = 5, beta = 1/2),
                    weight.f = as.weight(20, quadrature = TRUE), symmetric = TRUE)

# Old Scale weights with Gamma(5, 2) New Scale weights
unif_gamma <- plink(x, weights.t = as.weight(20),
                    weights.f = as.weight(20, quadrature = TRUE,
                    dist = 'gamma', alpha = 5, beta = 1/2), symmetric = TRUE)
norm_gamma <- plink(x, weights.t = as.weight(20, quadrature = TRUE),
                    weights.f = as.weight(20, quadrature = TRUE,
                    dist = 'gamma', alpha = 5, beta = 1/2), symmetric = TRUE)
gamma_gamma <- plink(x, weights.t = as.weight(20, quadrature = TRUE,
                     dist = 'gamma', alpha = 5, beta = 1/2),
                     weight.f = as.weight(20, quadrature = TRUE,
                     dist = 'gamma', alpha = 5, beta = 1/2), symmetric = TRUE)
```