Ability and Prior Distribution Mismatch: An Exploration of Common-Item Linking Methods

Applied Psychological Measurement
2017, Vol. 41(7) 545–560
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617707508
journals.sagepub.com/home/apm

Brandon LeBeau¹

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.

Keywords

calibration, linking, psychometrics, simulation

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 (1968) three-parameter logistic model:

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

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 , commonly set to 1.702 to transform the logistic ogive into a normal ogive (Camilli, 1994; De Ayala, 2013). A more thorough introduction to IRT can be found in De Ayala (2013) and Lord (1980).

---

¹University of Iowa, Iowa City, IA, USA

Corresponding Author:

Brandon LeBeau, University of Iowa, 311 Lindquist Center, Iowa City, IA 52242, USA.
Email: brandon-lebeau@uiowa.edu

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, commonly referred to as the “0, 1” ability scale. 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 operational 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 by S. Kim and Kolen (2007) has empirically studied the choice of the proficiency points and weights with characteristic curve methods.

This study aims to extend the work done by S. Kim and Kolen (2007) by empirically studying the impact of different specifications of proficiency points and weights with 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. Accurate estimation of 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):

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

These equations show the relationship between the ability scales, θ , and item parameters, a and b , for the old and new scales, represented by the o and n subscripts, respectively, as a function of the A and B linear transformation constants. The pseudo-guessing parameter remains unchanged in the old and new forms (i.e., $c_{jo} = c_{jn}$). The A and B transformation constants shown in Equation 1 are estimated from various linking methods discussed in 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). The nonequivalent group design is popular, in that respondents do not need to be statistically equivalent when given the assessment. The goal is to link the two groups onto a common scale after responses are obtained. More details on this and other linking and equating designs can be found in Kolen and Brennan (2004).

There are four common linking methods used with the common-item nonequivalent group design: mean/mean, mean/sigma, 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. Simulation studies have shown concurrent calibration to provide better estimates than linking methods in many scenarios and will be compared in this study (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. These two methods are different in how the A transformation constant is calculated. The mean/mean method uses a ratio of the mean of the estimated discrimination parameters, whereas the mean/sigma method uses a ratio of the standard deviation of the estimated difficulty parameters.

Although simple, it has been shown that these methods do not perform quite as well as the other two methods, commonly referred to as 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 and test characteristic curve for the Haebara (1980) and 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 (1980) method minimizes the following function:

$$H = H_1 + H_2, \quad (2)$$

where

$$H_1 = \sum_{ci}^{CI} \left[p\left(\theta_{jo} | \hat{a}_{jo}, \hat{b}_{jo}, \hat{c}_{jo}\right) - p\left(\theta_{jo} | \frac{\hat{a}_{jn}}{A}, A\hat{b}_{jn} + B, \hat{c}_{jn}\right) \right]^2 w_1(\theta_{jo}) \quad (3)$$

and

$$H_2 = \sum_{ci}^{CI} \left[p\left(\theta_{jn} | \hat{a}_{jn}, \hat{b}_{jn}, \hat{c}_{jn}\right) - p\left(\theta_{jn} | \hat{a}_{jo}A, \frac{\hat{b}_{jo} - B}{A}, \hat{c}_{jo}\right) \right]^2 w_2(\theta_{jn}), \quad (4)$$

where θ_{jo} and $w_1(\theta_{jo})$ are proficiency points and weights that represent the distribution of θ_{jo} (i.e., the old scale), and θ_{jn} and $w_2(\theta_{jn})$ are proficiency points and weights representing the distribution of θ_{jn} (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 described by Haebara. With the Stocking–Lord approach, the following is minimized to find the estimated A and B parameters:

$$S = S_1 + S_2, \quad (5)$$

where

$$S_1 = \left[\sum_{ci}^{CI} p(\theta_{jo} | \hat{a}_{jo}, \hat{b}_{jo}, \hat{c}_{jo}) - p\left(\theta_{jo} | \frac{\hat{a}_{jn}}{A}, A\hat{b}_{jn} + B, \hat{c}_{jn}\right) \right]^2 w_1(\theta_{jo}) \quad (6)$$

and

$$S_2 = \left[\sum_{ci}^{CI} p(\theta_{jn} | \hat{a}_{jn}, \hat{b}_{jn}, \hat{c}_{jn}) - p\left(\theta_{jn} | \hat{a}_{jo}A, \frac{\hat{b}_{jo} - B}{A}, \hat{c}_{jo}\right) \right]^2 w_2(\theta_{jn}). \quad (7)$$

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 using individual item characteristic curves. 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 in a second item.

Symmetric Linking

Symmetric linking takes into account information regarding the old and new scales, whereas unsymmetric linking 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 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 logic for the Stocking–Lord methods where S_1 and S_2 would be substituted for H_1 and H_2 in the discussion above (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 in the range of -4 to 4 from the standard normal distribution. With many options returned by common software used for linking, research on the effects of the prior distributions and weights on the transformation constant estimates is needed to guide practitioners.

To date, a single study by S. Kim and Kolen (2007) has explored the impact of distribution- and symmetry-related issues with characteristic curve methods. This study found that mismatch between the ability distributions and the prior distributions resulted in more linking error compared with 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.

This study looks to extend 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. Second, 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) will generalize across two software implementations. Finally, the primary outcome of interest is the bias of the transformation constants, and the population transformation constants were varied as part of the simulation design. This will allow for a greater extension 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 to the following three conditions: $A = 1$ and $B = 0$, $A = 1$ and $B = -1$, and $A = 1$ and $B = 1$.

The research questions for this study include the following:

Research Question 1: To what extent does the prior distribution have an impact on the estimation of the transformation constants?

Research Question 2: To what extent does symmetric linking have an impact on the estimation of the transformation constants?

Research Question 3: To what extent does the relationship between Questions 1 and 2 generalize across the simulation conditions specified?

Method

The following conditions were manipulated using a computer simulation, the ability distribution for Form Y (four levels: Normal (0, 1), $\gamma(5, 2)$, $\chi^2(4)$, $t(3)$), the percentage of common items (three levels: 25%, 35%, 45%), length of test (three levels: 25, 40, 55 items), and sample size for Y (two levels: 800, 1,500). The choice of ability distributions was made to include both symmetric (Normal and $t(3)$) and skewed distributions ($\gamma(5, 2)$ and $\chi^2(4)$) 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 new items are field tested, but there are only a small number of respondents or multiple administrations are not possible in field testing. In these situations, the number of new field-tested items may need to be larger compared with the number of common items. The simulation conditions discussed above were fully crossed in a factorial research design resulting in $4 \times 3 \times 3 \times 2 = 72$ conditions. The study used 1,000 replications for each simulation condition. Note that the ability distributions for Form Y were standardized based on the theoretical mean and standard deviation of each distribution. Finally, whenever the number of common items was a decimal (e.g., 8.75) it was rounded to the nearest 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 (three levels) and new scale (four levels). The three levels of the prior ability distribution for the old scale were 20 equally spaced proficiency points and weights between -4 and 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 characteristic curve methods. Note that the method moment linking methods and concurrent calibration will not vary for these within-study conditions. Therefore, there were a total of $72 \times 24 = 1,728$ characteristic study conditions, 72 concurrent calibration conditions, and 144 method moment study conditions for a total of 1,944 study conditions. A diagram highlighting the study conditions can be seen in the supplementary material.

In addition to these fully crossed simulation conditions, the transformation parameters representing the difference between the common-item parameters for Form X and Form Y were simulated from a uniform distribution. 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 .10, 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 six common items on a 25-item assessment, the first six items were common with Form X, the other 19 items were randomly simulated to represent new items. Form Y will then be linked onto the same scale as Form X. A diagram of the form simulation procedure can be seen in Figure 1.

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.

Analysis

After simulating the new responses for Form Y, item responses were calibrated with Bilog-MG (Mislevy & Bock, 1990). A three-parameter IRT model was used for all calibrations using software defaults. 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 with the known population constant values to assess the recovery of these transformed values, given the simulation conditions. In addition, concurrent calibration was used as an additional comparison. In these conditions, Form X and Form Y were calibrated in a single estimation procedure. The common items across the forms were estimated concurrently, and new parameter estimates were obtained for items not in common. The data structure looked as follows, where *CI* are common items, *F – X* are unique items on Form X, and *F – Y* are unique items on Form Y:

Figure 1. Diagram depicting simulation of item responses for the two forms.

$$\begin{pmatrix} CI_{1,1} & \dots & CI_{1,c} & F - X_{1,1} & \dots & F - X_{1,nx} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ CI_{p,1} & \dots & CI_{p,c} & F - X_{p,1} & \dots & F - X_{p,nx} \\ CI_{p+1,1} & \dots & CI_{p+1,c} & & & F - Y_{1,1} \dots F - Y_{1,ny} \\ \vdots & \ddots & \vdots & & & \vdots \\ CI_{p+q,1} & \dots & CI_{p+q,c} & & & F - Y_{q,1} \dots F - Y_{q,ny} \end{pmatrix}.$$

Form X was used as the reference group, and the mean and standard deviation of Form Y estimated via the concurrent calibration were used as estimates for the B and A transformation constants, respectively. A similar approach was used by S. Kim and Kolen (2007). An example Bilog-MG concurrent calibration control file is given in the supplemental material.

Bias of the transformation constants will be explored: $bias = \hat{\theta} - \theta$. The bias formula 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 linking methods (concurrent, Haebara, Stocking–Lord, mean/mean, and mean/sigma) and a second model that restricted the analysis to only 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 .001.

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

$$\text{RMSE} = \sqrt{\frac{\sum (\hat{\theta} - \theta)^2}{R}}, \quad (8)$$

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.

Results

The bias in the transformation constants was first explored descriptively where differences in the average bias across the ability distributions for the method moment, concurrent, and characteristic curve methods were found. On average, the bias in transformation constants tended to be much smaller for the $\gamma(5, 2)$ and normal distributions, and there was increased variation in estimates with the $\chi^2(4)$ and $t(3)$ distributions. In addition, there was evidence that the method moment methods had larger average bias statistics compared with the other methods. Little differences were observed among the other simulation factors such as sample size, number of common items, or length of test.

Finally, 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 or small relationships between the population A and B values ($r = .002$), the A population constants and the bias of the B constants ($r = -.02$), and the bias of the A and B constants ($r = -.07$). Moderate to strong correlations were found between the A population constants and the bias of the A constants ($r = -.26$), the B population constants and the bias of the B constants ($r = .20$), and the B population constants and the bias of the A constants ($r = -.42$). The correlations tended to be larger in many cases when looking at only the concurrent calibration results but were in the same direction. For example, the correlation between the B population constants and the bias of the A constants was $-.54$, and the correlation between the B population constants and the bias of the B constants was $.30$.

RMSE

Average RMSE values were calculated by the four ability distributions, five linking methods, and transformation population values. Figure 2 shows that the average RMSE values for most simulation conditions are close to 0. Increases in the average RMSE values occur as the B population transformation constant moves away from 0, particularly for the $\chi^2(4)$ ability distribution and characteristic curve methods. Concurrent calibration consistently had average RMSE values very close to 0 across all ability distributions and population transformation constant values.

The RMSE values for the estimated B transformation constants were similar to those for the estimated A transformation constants. Notably, there was evidence that characteristic curve methods had larger RMSE values as the population transformation constant values move away from no transformation (i.e., $A = 1$ and $B = 0$). This difference is most notable for extreme population B transformation constants (i.e., $B = \pm 2$). Concurrent calibration again, on average, had the smallest RMSE values.

Figure 2. Average RMSE of the estimated A transformation constant by three A population values, the B population values, ability distribution, and linking method.

Note. RMSE = root mean square error.

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 0 (i.e., 1 was subtracted from all A population values) to ensure that 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 .001 for all linking methods can be seen in the second and third columns of Table 1. Overall, the effect sizes shown in Table 1 explained 43% and 39% of the variation in the bias of the A and B transformation estimates, respectively. Sample size 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, linking method, and the interaction between these variables. Only a small amount of variation in the bias of the transformation constants was explained by the old and new prior distributions, less than 3% in all instances.

Average bias statistics by the linking method, ability distribution, and population transformation constants (only three levels of the A transformation constants are shown) can be seen in Figures 3 and 4 for A and B, respectively. These figures show that bias in the estimates of the transformation constants tends to occur as the population constants move away from no transformation. These differences are more pronounced in the method moment linking methods and in ability distributions that are not normally distributed. In particular, the mean/sigma method performance was poor across more simulation conditions compared with other methods.

Table 1. Effect Sizes From ANOVA Model Exploring Absolute Bias of A and B Parameters for Characteristic Curve and Method Moment Linking Methods.

Variable	Con & MM & CC		CC only	
	$\hat{\eta}^2$ A	$\hat{\eta}^2$ B	$\hat{\eta}^2$ A	$\hat{\eta}^2$ B
Method	.037	.039		
Ability dist.	.044	.023	.064	.034
Prior dist. old	.022	.028	.040	.041
Prior dist. new	.007	.017	.013	.025
A Pop.	.050		.068	
B Pop.	.128	.046	.175	.038
Num items		.003		.004
Method:Ability dist.	.003	.002		
Method:Num items	.003			
Method:A pop.	.005	.011		
Method:B pop.		.004		
Ability dist.:Prior dist. old	.007	.009	.013	.013
Ability dist.:Prior dist. new	.004	.006	.007	.009
Ability dist.:A pop.	.003	.003	.004	.003
Ability dist.:B pop.	.074	.051	.094	.058
Prior dist. old:Prior dist. new	.008	.024	.014	.035
Prior dist. old:A pop.	.005	.004	.009	.006
Prior dist. old:B pop.		.005		.008
Prior dist. new:A pop.	.002	.002	.003	.003
Prior dist. new:B pop.	.004	.008	.008	.011
Num items:B pop.	.002		.004	.002
A pop.:B pop.	.003	.063	.003	.077
Method:Ability dist.:B pop.	.010	.002		
Ability dist.:Prior dist. old:Prior dist. new	.002	.008	.004	.012
Ability dist.:Prior dist. old:B pop	.003	.005	.005	.007
Ability dist.:Prior dist. new:B pop	.003	.008	.005	.011
Ability dist.:A pop.:B pop.	.003	.004	.005	.005
Prior dist. old:Prior dist. new:A pop	.002	.003	.003	.004
Prior dist. old:Prior dist. new:B pop		.012	.002	.017

Note. Con = concurrent; MM = method moment; CC = characteristic curve; ":" = interaction; old = old scale; new = new scale; num = number.

Concurrent calibration again performed very well in many conditions, as shown in Figures 3 and 4; however, there are conditions in which characteristic curve methods on average outperform concurrent calibration. For example, characteristic curve methods tend to outperform concurrent calibration when the population B transformation constants are closer to 2 in absolute value when the ability distribution was normal or $t(3)$. In comparison, concurrent calibration did perform better on average in the extreme values when the ability distribution was $\chi^2(4)$ and $\gamma(5, 2)$. When the population B transformation constant was close to 0, concurrent calibration and characteristic curve methods showed similar performance.

Characteristic Curve Methods Only

A second set of models were run that omitted the method moment and concurrent linking methods and only analyzed characteristic curve methods. The effect sizes larger than .001 can be seen in the last two columns of Table 1, and these explained approximately 54% and 42% of

Figure 3. Average bias of the estimated A transformation constant by three A population values, B population values, ability distribution, and linking method.

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 characteristic curve methods, suggesting no meaningful difference between the estimates from Haebara and Stocking-Lord methods. The prior distributions, old and new scales, explained small amounts of variations, less than 3% for both the A and B transformation constants. The range of average bias for the A and B transformation constants using characteristic curve methods was smaller than $\pm .5$ under most conditions.

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 ($\chi^2(4)$) or has heavy tails ($t(3)$) the amount of bias increases, particularly for larger B population constant values. In addition, the prior distribution for the old scale has a very small effect on the bias; however, the $\gamma(5, 2)$ and uniform distribution performed worse compared with the normal distribution, even when this matches the ability distribution.

Empirical Example

An empirical data example is used to show the differences in the different linking methods while specifying different proficiency points and weights. Data from nonequivalent 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 perform these linking computations with the plink R package can be seen in the supplemental material.

Twelve different combinations of weights were used on the empirical data to show the differences in estimates of 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 nonsymmetric linking designs. To explore differences when symmetry

Figure 4. Average bias of the estimated B transformation constant by three A population values, B population values, ability distribution, and linking method.

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. Grade 3 was used as the base group (i.e., IRT parameters were placed on the Grade 3 scale).

The differences in the linking specifications were modest within each linking method and prior distribution for the old scale; however, small differences were associated with the method. Particularly, the Haebara linking constants tend to be slightly smaller in many conditions compared with the Stocking–Lord constants. These differences are very modest and are likely not of practical importance. In addition, the $\gamma(5, 2)$ distribution for the proficiency points and weights for the old scale produces smaller estimates going from Grades 7 to 8. The linking constant estimates can be seen in Table 2.

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 devoted to nonequivalent groups using 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) has explored the implications for different specifications of these prior distributions of the old and new scales. 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 in an attempt to improve generalizability.

The results showed that the prior distribution for the old and new scales is not a strong predictor when looking at bias of the transformation constants, and the results generalize across the simulation conditions considered in this study. For example, the effect due to prior distribution

Table 2. Linking Constant Values by Linking Method and Prior Distribution for Old and New Scales by Grade.

Method	Prior distribution for old scale	Prior distribution for new scale	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.

was similar across the number of common items, sample size, and linking method. If the transformation constants differ significantly from no transformation, there is evidence that using a normal prior distribution does help limit bias; however, the bias in the other conditions is relatively small in most conditions.

These results provide reassurances to researchers or practitioners that the results are not sensitive to the prior distribution, and even provide evidence that using simple prior distributions, such as a normal distribution, provides 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 characteristic curve methods is robust with regard to the prior distributions. Similar to prior work, method moment methods showed larger amounts of bias compared with characteristic curve methods (Hanson & Béguin, 2002; S. Kim & Lee, 2006), suggesting that the method moment methods are less desirable than characteristic curve methods. With the improvements in computers, the computations needed to perform characteristic curve linking methods are quick to perform and should be used over method moment methods.

Prior work (Hanson & Béguin, 2002; S. Kim & Kolen, 2007) has shown that concurrent calibration is superior to characteristic curve methods. This result was replicated in many simulation conditions considered in the current study, namely when transformation constant values are not too extreme or with skewed distributions (e.g., $\chi^2(4)$ and $\gamma(5, 2)$). However, evidence was found that when the transformation constants are larger, indicating a larger difference in the average ability level or variance of the nonequivalent groups, characteristic curve methods can outperform concurrent calibration. Characteristic curve methods also performed better when the ability distribution has heavy tails (e.g., $t(3)$). This suggests that a careful exploration of the ability distribution is needed to have a sense of the method that may produce the least amount of bias. If there is evidence that the distribution is skewed, concurrent calibration may perform better than other methods. However, if there is evidence that the distribution has heavy tails or the distributions of the two groups differ significantly, characteristic curve methods may be a better choice.

The current study only used the three-parameter IRT model; therefore, the results pertain only to this model choice. Simpler models that do not estimate the guessing (lower asymptote) parameter may offer better performance due to more accurate estimation of the other model parameters, namely the discrimination and difficulty.

Finally, the number of common items did not significantly affect the estimates of the A and B transformation constants. The smallest number of common items considered in this study was six items, which suggests that on average there is no difference in the average bias when using six common items or 10. 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 six common items compared with larger number of common items. This is not surprising, given that the estimate of the standard deviation of the item difficulties would be based on only a few items leading to more variable estimates.

Even though the number of common items did not significantly affect estimates of the A and B transformation constants, practical considerations will affect this choice. For example, the common items should be representative of the entire assessment in terms of both difficulty and content. If new items were field tested to appear on a mathematics assessment, the common items may need to span subgroups such as algebra and geometry. In addition, within each of

those subgroups, items spanning the Difficulty scale would want to be included. This can quickly increase the number of common items needed to ensure that accurate linking constants are estimated, but only including a few items within these subgroups may be needed for sufficient estimation of the transformation constants.

Conclusion

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, care needs to be taken when linking these two forms. The choice between concurrent calibration and linking via characteristic curve methods depends in part on the shape and magnitude of difference in the ability distributions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online supplementary material are available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617707508>.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. (pp. 395-479). Reading, MA: Addison-Wesley.
- 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, 293-295.
- De Ayala, R. J. (2013). *Theory and practice of item response theory*. New York, NY: Guilford Press.
- Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, 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, 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, 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, 371-397.

- Kim, S., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*, 53-76.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
- Mislevy, R. J., & Bock, R. D. (1990). *Bilog 3: Item analysis and test scoring with binary logistic models*. Skokie, IL: Scientific Software International.
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org>
- Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*, 137-167.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 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*, 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/>