Factors affecting the variability of IRT equating coefficients

Factors Affecting the Variability of IRT Equating Coefficients

Michela Battauz

michela.battauz@uniud.it

Department of Economics and Statistics, University of Udine,
via Tomadini 30/A, 33100 Udine, Italy

Knowing the effect of the factors that can influence the variability of the equating coefficients is an important tool for the development of the linkage plans. This paper explores the effect of various factors on the variability of IRT equating coefficients. The factors studied are the sample size, the number of common items, the length of the chain and the possibility of averaging the equating transformations related to different paths that connect the same two forms. Both asymptotic and simulations results are provided.

Keywords and Phrases: accuracy, bisector, double linking, equating, multiple linking, Rasch, standard errors.

1 Introduction

Equating is a process that permits the comparison of scores obtained on different test forms. In this paper, item response theory (IRT) methods for test equating will be considered under the common-item nonequivalent groups design (Kolen & Brennan, 2004).

IRT equating methods provide a linear transformation of person and item parameters and the coefficients of this function are called equating coefficients. The equating coefficients are subject to sampling variation because they are estimated on the basis of the item parameter estimates. In order to obtain reliable equatings, it is then important to limit the variability of the equating coefficients.

Several factors are expected to influence the variability of the equating coefficients. Considering just two forms, some factors are the sample size, the number of common items or the equating method chosen, where the sample size is defined as the number of examinees used for the calibration of the item parameters in a test. However, many testing programs equate test forms across several administrations, thus introducing complex linkage plans that include chains and the connection of forms through different paths. Two further factors that can influence the variability of the equating coefficients are then the length of the chain linking two forms and the opportunity of averaging the scale conversions deriving from different paths. Knowing the impact of these factors on the variability of the equating coefficients is important in order to design the linkage plan.

In recent years, asymptotic standard errors of the equating coefficients were derived (Ogasawara, 2000, 2001; Battauz, 2013b). These works contain simulation results that provide some insight on the factors that affect the variability of the equating coefficients. However, analytic results on the factors that have an impact on the variance of the equating coefficients and simulation studies that investigate in a systematic manner the effect of these factors are missing in the literature.

This paper investigates the effect of various factors on the variability of equating coefficients, providing both asymptotic and simulation results. The factors studied are the sample size, the number of common item, the length of the chain, and averaging the transformations obtained from different paths. Asymptotic results apply to any IRT model or equating method used. In simulations, the IRT models considered are the

Rasch and the two-parameter logistic models, while the equating methods used are the mean-mean (Loyd & Hoover, 1980) and the Haebara (Haebara, 1980) methods.

The paper is structured as follows. Section 2 summarizes IRT equating methods, Section 3 provides some asymptotic results on the impact of the factors mentioned on the standard deviation of the equating coefficients and Section 4 contains the results of several simulation studies. Finally, a discussion is given in Section 5.

2 IRT test equating

In the three-parameter logistic model (van der Linden & Hambleton, 1997), the probability of a positive response on item j for a person with ability θ is given by

$$p_j(\theta; a_j, b_j, c_j) = c_j + (1 - c_j)\frac{\exp\{Da_j(\theta - b_j)\}}{1 + \exp\{Da_j(\theta - b_j)\}}, \tag{1}$$

where a_j is the item discrimination parameter, b_j is the item difficulty parameter, c_j is the item guessing parameter and D is a constant typically set to 1.7. Setting the guessing parameters to zero yields the two-parameter logistic (2PL) model. The Rasch model is obtained by also fixing the discrimination parameters to one.

When different forms are not administered to the same population, item parameter and person ability estimates are not comparable because they are expressed on different measurement scales. The equating process permits the comparison of estimates obtained from different populations by expressing them on the same measurement scale. The following equations permit the transformation of parameters of form $g - 1$ to the scale of form g

$$\theta_g = A_{g-1,g}\,\theta_{g-1} + B_{g-1,g}, \quad a_g = \frac{a_{g-1}}{A_{g-1,g}} \quad \text{and} \quad b_g = A_{g-1,g}\,b_{g-1} + B_{g-1,g}, \tag{2}$$

where $A_{g-1,g}$ and $B_{g-1,g}$ are the equating coefficients. These coefficients can be estimated by using moments of item parameters (Kolen & Brennan, 2004, §6.3.2), or response function methods (Kolen & Brennan, 2004, §6.3.3).

Suppose that two forms are linked through a chain of forms that present common items in pairs. Define the path from form 1 to form l as $p = \{1, \ldots, l\}$. According to Battauz (2013b), the equating coefficients transforming the scale of θ_1 to that of θ_l are given by

$$A_{(p)} = A_{1,\ldots,l} = \prod_{g=2}^{l} A_{g-1,g} \qquad \text{and} \qquad B_{(p)} = B_{1,\ldots,l} = \sum_{g=2}^{l} B_{g-1,g}\, A_{g,\ldots,l}\,, \tag{3}$$

where $A_{g,\ldots,l} = \prod_{h=g+1}^{l} A_{h-1,\,h}$ is the coefficient that links form g to form l.

When two forms are linked through more than one path, scale conversions can be averaged in order to obtain a single equating relationship. To this end, the symmetry property, which requires that the inverse function of the average equating function equals the average of the inverse functions, is a desirable property. However, the mean does not satisfies this property that is instead satisfied by the bisector method (Holland & Strawderman, 2011; Battauz, 2013b). The bisector method yields a weighted average of the equating coefficients

$$A_{1l}^{*} = \sum_{p=1}^{P} A_{(p)} w_p \qquad \text{and} \qquad B_{1l}^{*} = \sum_{p=1}^{P} B_{(p)} w_p, \tag{4}$$

where

$$w_p = \frac{n_p (1 + A_{(p)}^2)^{-1/2}}{\sum_{b=1}^{P} n_b (1 + A_{(b)}^2)^{-1/2}}, \tag{5}$$

P is the number of paths that link forms 1 and l, and n_p are optional weights associated with each path. Note that, for the Rasch model, the bisector method is equivalent to the (weighted) mean because the coefficients $A_{(p)}$ are all equal to 1.

Standard errors of equating coefficients can be computed by using the delta method. See Ogasawara (2000) and Ogasawara (2001) for direct equating coefficients and Battauz

(2013b) for chain and bisector equating coefficients.

3 Asymptotic results

In this section some asymptotic results on the influence of the factors affecting the variability of the equating coefficients will be derived. Only the equating coefficient B will be considered in this section, because analogous results can be derived for the equating coefficient A. The sample size will be denoted by n and the number of common items will be denoted by m. Hence, the number of item parameters that should be equated between two forms is $2m$, as they include both discrimination and difficulty parameters. For the sake of simplicity, n and m will be assumed constant across forms, and the length of the chains, l, will be assumed constant for different paths. Let $\boldsymbol{\alpha}_g$ be the vector of item parameters of form g. When item parameters are estimated by using the marginal maximum likelihood method (Bock & Aitkin, 1981), standard likelihood theory can be applied. Then, the elements of the asymptotic variance-covariance matrix of the estimate $\hat{\boldsymbol{\alpha}}_g$, are of order $O(n^{-1})$.

Consider a chain of forms related to path $p = \{1, \ldots, l\}$. Let $\boldsymbol{\alpha}_{(p)} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_l^\top)^\top$ be the vector containing all the item parameters related to the forms that compose the path and $\text{acov}(\hat{\boldsymbol{\alpha}}_{(p)})$ the asymptotic variance-covariance matrix of the estimate $\hat{\boldsymbol{\alpha}}_{(p)}$. The j-th item parameter of form g will be denoted by α_{gj} The order of the partial derivatives of the chain equating coefficients with respect to the item parameters, $\frac{\partial B_{(p)}}{\partial \alpha_{gj}}$, can be derived from the formulas given in Ogasawara (2000), Ogasawara (2001) and Battauz (2013b), and they are either of order $O(m^{-1})$ or zero. Then, the asymptotic order of the variance of the chain equating coefficient is

$$
\begin{aligned}
\mathrm{avar}(\hat{B}_{(p)}) &= \frac{\partial B_{(p)}}{\partial \boldsymbol{\alpha}_{(p)}^{\top}} \,\mathrm{acov}(\hat{\boldsymbol{\alpha}}_{(p)})\, \frac{\partial B_{(p)}}{\partial \boldsymbol{\alpha}_{(p)}} && (6)\\
&= \sum_{g=1}^{l}\sum_{j=1}^{2m}\left(\frac{\partial B_{(p)}}{\partial \alpha_{gj}}\right)^{2}\mathrm{avar}(\hat{\alpha}_{gj}) + \sum_{g=1}^{l}\sum_{j=1}^{2m}\sum_{\substack{k=1\\k\neq j}}^{2m}\frac{\partial B_{(p)}}{\partial \alpha_{gj}}\frac{\partial B_{(p)}}{\partial \alpha_{gk}}\mathrm{acov}(\hat{\alpha}_{gj},\hat{\alpha}_{gk}) && (7)\\
&= O(l)O(m^{-1})O(n^{-1}) + O(l)O(n^{-1}) && (8)\\
&= O(l)O(n^{-1}). && (9)
\end{aligned}
$$

This result shows that the order of the variance of the equating coefficient increases with the length of the chain and decreases with the sample size. Instead, the number of common items does not decrease the asymptotic order of the variance of the equating coefficient. When m tends to infinity, just the first term in equation (7) tends to zero. So, the asymptotic variance of the equating coefficient tends to the second term in equation (7), which is a weighted mean of the covariances. However, since the covariances are smaller than the product of the corresponding standard deviations, the variance of the equating coefficient is expected to diminish when m augments. This point will be further investigated by simulation studies in the next section.

The asymptotic order of the variance of direct equating coefficients are a special case of chain equating coefficients with $l = 2$.

In case of average equating coefficients, let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{(p)})_{p=1,\ldots,P}$ be the vector containing all the item parameters entering in at least one of the paths averaged, and $\mathrm{acov}(\hat{\boldsymbol{\alpha}})$ the asymptotic variance-covariance matrix of the estimate $\hat{\boldsymbol{\alpha}}$. The j-th item parameter of the g-th form in path p will be denoted by $\alpha_{(p)gj}$. From the formulas given in Battauz (2013b), it can be derived that the derivatives $\frac{\partial B_{1l}^{*}}{\alpha_{(p)gj}}$ are either of order $O(P^{-1}m^{-1})$ or zero, provided that w_p is of order $O(P^{-1})$. The asymptotic variance of the average equating coefficient is then given by

$$
\begin{aligned}
\mathrm{avar}(\hat{B}^*_{1l}) &= \frac{\partial B^*_{1l}}{\partial \boldsymbol{\alpha}^\top}\,\mathrm{acov}(\hat{\boldsymbol{\alpha}})\,\frac{\partial B^*_{1l}}{\partial \boldsymbol{\alpha}} &(10)\\
&= \sum_{p=1}^{P}\sum_{g=1}^{l}\sum_{j=1}^{2m}\left(\frac{\partial B^*_{1l}}{\partial \alpha_{(p)gj}}\right)^2 \mathrm{avar}(\hat{\alpha}_{(p)gj})+ \\
&\qquad \sum_{p=1}^{P}\sum_{b=1}^{P}\sum_{g=1}^{l}\sum_{h=1}^{l}\sum_{j=1}^{2m}\sum_{\substack{k=1\\k\neq j}}^{2m}\frac{\partial B^*_{1l}}{\partial \alpha_{(p)gj}}\frac{\partial B^*_{1l}}{\partial \alpha_{(b)hk}}\mathrm{acov}(\hat{\alpha}_{(p)gj},\hat{\alpha}_{(b)hk}) &(11)\\
&= O(P^{-1})O(l)O(m^{-1})O(n^{-1})+O(l)O(n^{-1}) &(12)\\
&= O(l)O(n^{-1}). &(13)
\end{aligned}
$$

The second term of equation (11) of is of order $O(l)O(n^{-1})$ and not $O(l^2)O(n^{-1})$ because $\mathrm{acov}(\hat{\alpha}_{(p)gj},\hat{\alpha}_{(b)hk})$ is not zero only when form g of path p and form h of path b are the same. In most cases, two paths share only the first and the last form. Analogously to the previous case, the order of the variance is affected by the length of the chain and the sample size, while the number of common items does not affect the order of the variance. The effect of the number of paths averaged is similar to the effect of the number of common items. Increasing the number of paths averaged does not affect the asymptotic order of the average equating coefficient. When P tends to infinity the variance tends to the second term in equation (11), that is a weighted mean of the covariances of the item parameters. However, since the covariances are smaller than the product of the corresponding standard deviations, the variance of the equating coefficient is expect to diminish for increasing values of m and P. The reduction will be stronger when the correlation between the item parameter estimates is small.

4 Simulation results

The variability of the equating coefficients was also investigated by means of several simulations studies. The factors that were considered are the sample size, the number of common items, the length of the chain, the equating method used, and averaging the equating coefficients by using the bisector method. Three different linkage plans were designed to understand the effect of these factors on the variability of the equating coefficients and they are represented in Figure 1.

[Figure 1 about here.]

Data were generated according to the Rasch and the 2PL models. The sample size considered for each form was n=250, 500, 1000, 2000, 4000 and 8000, and was taken constant across forms in the same simulation study. The number of common items were 5, 10, 15 or 20, while the length of the chain linking two forms varies from 2 to 10. The equating methods used were the mean-mean and the Haebara methods. Since the two methods gave very similar results (the lowest value of the correlation coefficient of the equating coefficients estimates is 0.96), only the results obtained with the mean-mean method are shown here. For each setting, 500 data sets were generated. Data were generated using the R software (R Development Core Team, 2013). The item parameters were estimated using the R package `ltm` (Rizopoulos, 2006) and the equating coefficients were computed using the R package `equateIRT` (Battauz, 2013a). In the following, the results obtained for each linkage plan will be presented.

Results for the linkage plan 1. The linkage plan 1 (Figure 1a) is intended to study the effect of all factors mentioned on the variability of the equating coefficients for a single chain. There are 10 forms and each form is composed by 40 items and presents 5, 10, 15 or 20 items in common with adjacent forms. Person ability parameters were generated from a normal distribution with mean equal to -0.25 for odd forms and 0.25 for even forms. The standard deviation was taken equal to 1 for the Rasch model, while for the 2PL model the standard deviation was set to 1 for odd forms and 1.2 for even forms. In order to obtain items with difficulties aligned with person abilities, item parameters were assigned equispaced values in the range obtained by adding and subtracting 0.5 from the mean of person parameters. The discrimination parameters, in the 2PL model, were generated from the uniform distribution with range [0.8, 1.2]. Form 1 was equated to forms 2, 4, 6, 8 and 10 to produce equatings with different lengths of the chain.

Figures 2 and 3 represent the standard deviation of the estimated B equating coefficient for the Rasch model. Figure 2 shows that increasing the number of common items the standard deviation of the estimated B equating coefficient decreases and that the reduction is larger when the number of common items is small. Figure 3 shows that the length of the chain is positively correlated with the standard deviation of the equating coefficient. Nevertheless, both figures show that the stronger effect can be imputed to the sample size. In fact, the lowest values of the standard deviation of the equating coefficient can be achieved only when the sample size is large. Furthermore, with large samples the effect of the number of common items and the length of the chain is strongly reduced.

The effects of these factors are consistent with the findings of Section 3, confirming that increasing the number of common items and the sample size reduces the variability of the equating coefficients, while augmenting the lenght of the chain yield larger variability. Simulations confirm also that the number of common items has a limited effect on the standard deviation of the equating coefficients and that increasing this factor does not reduce the variability to zero.

[Figure 2 about here.]

[Figure 3 about here.]

To better understand the influence of these factors on the variability of the estimated

B coefficient, some regression models were estimated. The dependent variable is the standard deviation of the equating coefficient, while the independent variables were the number of common items and the length of the chain. A different regression model was fitted for each sample size. Despite the number of observations for each regression is small, and equals to 20, the analysis comply with the purpose of quantifying the average effect of each factor. The number of common items was transformed using the inverse function in order to better represent the behavior of this factor. In order to make the results more meaningful, the length of the chain was centered at 2, so that the intercept represents the standard deviation of the equating coefficient when m tends to infinity and $l = 2$. Table 1 shows the regression coefficients estimates. The results of the regression models confirm that the sample size presents the biggest influence and that the effect of the other factors disappears for very large samples. For example, the effect of the number of common items, m, when the chain length, l, is equal to 2, is given by $0.2059/m$ for $n = 250$. This means that, keeping $l = 2$, when $m = 1$ the standard deviation is expected equal to $0.0938 + 0.2059 = 0.2997$, when $m = 2$ the standard deviation is $0.0938 + 0.2059/2 = 0.1968$, when $m = 3$ the standard deviation is $0.0938 + 0.2059/3 = 0.1624$ and so on. Increasing the sample size from 200 to 8000 this effect is reduced, varying from 0.2059 to 0.0385. The effect of l is instead linear. This means that, for m tending to infinity, every form added to a chain leads to an average increment of 0.0039 to the standard deviation, when $n = 250$. Increasing the sample size to 8000, this effect is gradually reduced from 0.0039 to 0.0009. The positive value of the interaction between $1/m$ and $l - 2$ indicates that the effect of m is higher for larger values of l and that the effect of l is higher for smaller values of m. Consider for example the case $n = 250$. The effect of m is $(0.2059 + 0.0833 \cdot (l - 2))/m$, while the effect of l is $(0.0039 + 0.0833/m) \cdot (l - 2)$. The lowest value of the coefficients of determination of the regression models is 0.97, indicating a very good fit, though this value should be considered taking into account the small number of observations.

[Table 1 about here.]

Figures 4, 5, 6 and 7 represent the standard deviation of the A and B equating coefficients for the 2PL model with respect to the number of common items and the length of the chain. The results of the regression models, analogous to those fitted for the previous case, are given in Tables 2a and 2b. The graphs and the regression model estimates show that the influence of the factors studied on the standard deviation of the equating coefficients for a 2PL model is very similar to what emerged for the Rasch model. In particular, the regression coefficients estimated for the B equating coefficient are very similar to those obtained for the B equating coefficient for the Rasch model. Comparing the coefficients obtained for the A and B equating coefficients, the intercept is smaller while the other parameters tend to be greater for coefficient A. This indicates that the effect of the number of common items and the length of the chain is slightly higher for the A equating coefficient, thus permitting to reach lower values of the standard deviation. Anyway, these findings confirm that the number of common items and the sample size present a negative effect on the standard deviation of the equating coefficients, while the length of the chain presents a positive effect. The effect of n and m is stronger for smaller values of them, while l presents a linear effect. The effect of all these factors is attenuated when the standard deviation of the equating coefficient is reduced by the effect of another factor.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Table 2 about here.]

Results for the linkage plans 2 and 3. The linkage plans 2 and 3 (Figure 1b and 1c) are designed to observe the effect of averaging scale conversions obtained from different paths.

Relative to the second linkage plan, the person parameters were generated from a normal distribution with mean equal to -0.25 for form 1, equal to 0 for forms from 2 to 5, and equal to 0.25 for form 6. The standard deviation was equal to 1 for every form for the Rasch model, while for the 2PL model the standard deviation was equal to 1 for forms from 1 to 5 and equal to 1.2 for form 6. Each form is composed by 80 items and item parameters were generated with the same mechanism used in the previous simulation study. The parameters of form 1 were converted to the scale of form 6 through the four different paths going through forms 2, 3, 4 or 5. Average equating coefficients were obtained using the bisector method. In this case, weighting the paths as proposed in Battauz (2013b) gives the same results of the unweighted bisector because the various paths are perfectly symmetric. For this reason only the results obtained with the unweighted bisector are reported.

Figure 8 regards the Rasch model. On the top of the figure we can find a plot of the standard deviations of the B equating coefficient with 5 common items for path $\{1, 2, 6\}$ against the standard deviations of the B coefficient with 10, 15 and 20 common items for the same path. On the bottom of the figure the standard deviations of a single path are plotted against the standard deviations of the average B coefficient obtained by using 2, 3 or 4 paths. For each plot there are 6 points, relative to the various sample sizes. For each plot, a regression model with intercept forced to zero was fitted. The regression line is drown on the plot and the estimated coefficient $\hat{\beta}$ is reported.

Comparing the regression coefficients on the top of the figure with those on the bottom of the figure, it is possible to observe very similar values. This shows that increasing the number of common items in a given chain, or averaging the equating coefficients obtained from different chains, yield very similar results in terms of reduction of the standard deviation of the equating coefficient, thus confirming the results presented in Section 3. In fact, the asymptotic results obtained showed that the number of common items and the number of chains averaged have the same effect on the order of the asymptotic variance of the equating coefficients.

Figure 9 represents the regression coefficients estimates obtained by regressing the standard deviation of the average equating coefficient on the standard deviation of a single equating coefficient. The regression coefficients are plotted against the number of common items and different lines refer to the number of chains averaged. The values represented for 5 common items are those reported on Figure 8. The goodness of fit was very good as indicated by the coefficients of determination that were all very close to one. The figure shows that the relative reduction of the standard deviations is larger for smaller values of the common items. It shows also that increasing the number of chains averaged produces smaller standard deviations, but this effect is attenuated when the number of common items is high.

[Figure 8 about here.]

[Figure 9 about here.]

Figures 10 and 11 represent the plots of the standard deviations of the A and B equating coefficients for the 2PL model. Analogously to Figure 8, on the top of the figures the standard deviations for a single path obtained with 5 common items are plotted against the standard deviations with 10, 15 and 20 common items. On the bottom of the figures the standard deviations for a single path obtained with 5 common items are plotted against the standard deviations of the average coefficients using 2, 3 or 4 chains and 5 common items. The regression lines with zero intercept are drown on the figures and the regression coefficient is reported. The figures show that the reduction of the standard deviation of the equating coefficients obtained by adding common items or by averaging the coefficients from different paths is extremely similar. Furthermore, it is possible to observe that the equating coefficient A presents a larger gain with respect to the equating coefficient B.

The regression coefficients reported on the bottom of Figures 10 and 11 for the average coefficients are represented in Figure 12, together with those obtained for 10, 15 and 20 common items. These figures show that coefficient A presents a larger reduction and that the gain is attenuated for greater values of the common items.

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

In the third linkage plan, form 3 is connected to form 1 directly or using a chain going through form 2. So, the parameters of form 1 were converted to the scale of form 3 using the direct and the indirect links. The equating coefficients obtained were then averaged using the bisector and the weighted bisector method. Weights were determined as proposed in Battauz (2013b). In this linkage plan, every form is composed by 40 items and shares 5, 10, 15 or 20 items in common with all other forms. Person parameters were generated from a normal distribution with means equal to -0.25, 0 and 0.25 for forms from 1 to 3 and standard deviations equal to 1 for the Rasch model and equal to 1, 1 and 1.2 for the 2PL model. Figures 13 and 14 show the results obtained. These figures represent the regression coefficients obtained by regressing the standard deviations of average equating coefficients on the standard deviations of the direct equating coefficient. For each regression, the intercept was forced to zero and the fitting was extremely good (the coefficients of determination were very close to one). It is possible to observe that the B equating coefficient presents very similar results for the Rasch model (Figure 13) and the 2PL model (right panel of Figure 14). Consistently with the previous findings, the A equating coefficient presents a larger gain (left panel of Figure 14). All figures show that the weighted bisector produces an higher reduction, though the difference is not very important. This is due to the fact that the two paths that were averaged do not present equating coefficients with very different variability.

[Figure 13 about here.]

[Figure 14 about here.]

Comparing the second and the third linkage plans, it is possible to observe that the relative reduction of the standard deviation of the equating coefficients is smaller for the latter. This is due to the fact that the relative reduction is computed with respect to the direct link in the third linkage plan. The first linkage plan already showed that for direct links ($l = 2$) the reduction that can be attained is small (see for example Figure 2). Furthermore, in the third linkage plan the gain is even smaller because the direct link was averaged with a chain of length 3, that presents larger variability than the direct link.

5 Discussion

This paper provides asymptotic and simulations results on the effect of the factors that can influence the variability of the IRT equating coefficients. These findings constitute a useful information for the development of linkage plans, so that the sampling variability can be controlled together with the other practical issues that should be taken into account.

This work focuses on the standard deviation of the equating coefficients and does not consider the bias. An alternative option would have been the determination of the mean square error of the equating coefficients. Since the equating coefficients obtained with the mean-mean and the Haebara methods are nearly unbiased (see for example Ogasawara, 2001), in this paper it was preferred to show only the results on variability.

According to the findings of this article, in order to obtain very precise equating coefficients, it is necessary to have large samples. With small samples, the number of common items and the length of the chain can assume a substantial role. The effect of adding a new path to link two forms and averaging the scale conversions was found similar to the effect of adding further common items to an existing path. Anyway, both these factors present a limited effect and the increment of them cannot make the standard deviation of the equating coefficients tend to zero. However, the simulations shown here do not account for departures from model assumptions. For example, the effect of item parameter drift was studied in Wells et al. (2002), who found that increasing the percentage of drifting items yields larger differences in person parameter estimates. In this respect, having a larger number of common items can attenuate the effect of fluctuations of a subset of item parameters. Further simulation studies, not presented here, show that the same effect can be achieved by averaging the scale conversions of different paths. In fact, the effect of item parameters drift in one path can be attenuated by the information provided by other paths that link the same forms. Adding further connections between two forms can be favorable because some cause of item parameter drift, like security breaches or changes in examinees demographics, are likely to interest more than one item in the same administration, while a new path could be unaffected by these problems. In conclusion, although the number of common items has a limited effect on the variability of equating coefficients, it is fundamental to maintain a certain number of common items to ensure robustness to departures from assumptions. Furthermore, it should be taken in consideration the fact that the number of common items have an important effect on the variability of the equating coefficients when the sample size is small, especially with long chains.

References

Battauz, M. (2013a). *equateIRT: Direct, chain and average equating coefficients with standard errors using IRT methods.* R package version 1.0-1.

Battauz, M. (2013b). IRT test equating in complex linkage plans. *Psychometrika*, 78, 464–480.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.

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

Holland, P. W. & Strawderman, W. E. (2011). How to average equating functions if you must. In von Davier A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 89–107). New York: Springer-Verlag.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (Second Edition).* New York: Springer-Verlag.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the rasch model. *Journal of Educational Measurement*, 17, 179–193.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce)*, 51, 1–23.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement*, 25, 53–67.

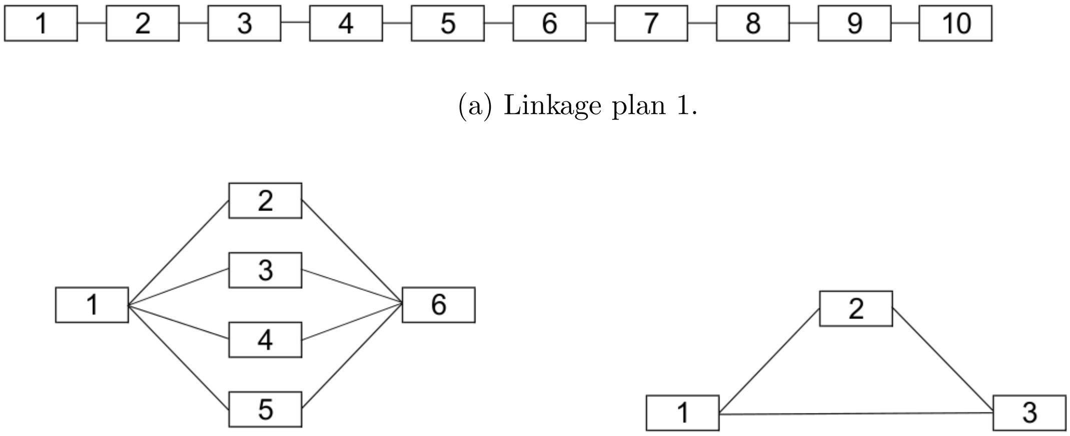
R Development Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17, 1–25.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.

Wells, C. S., Subkoviak, M. J., & Serlin, R. S. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77–87.

Figure 1: Linkage plans of the simulation studies.

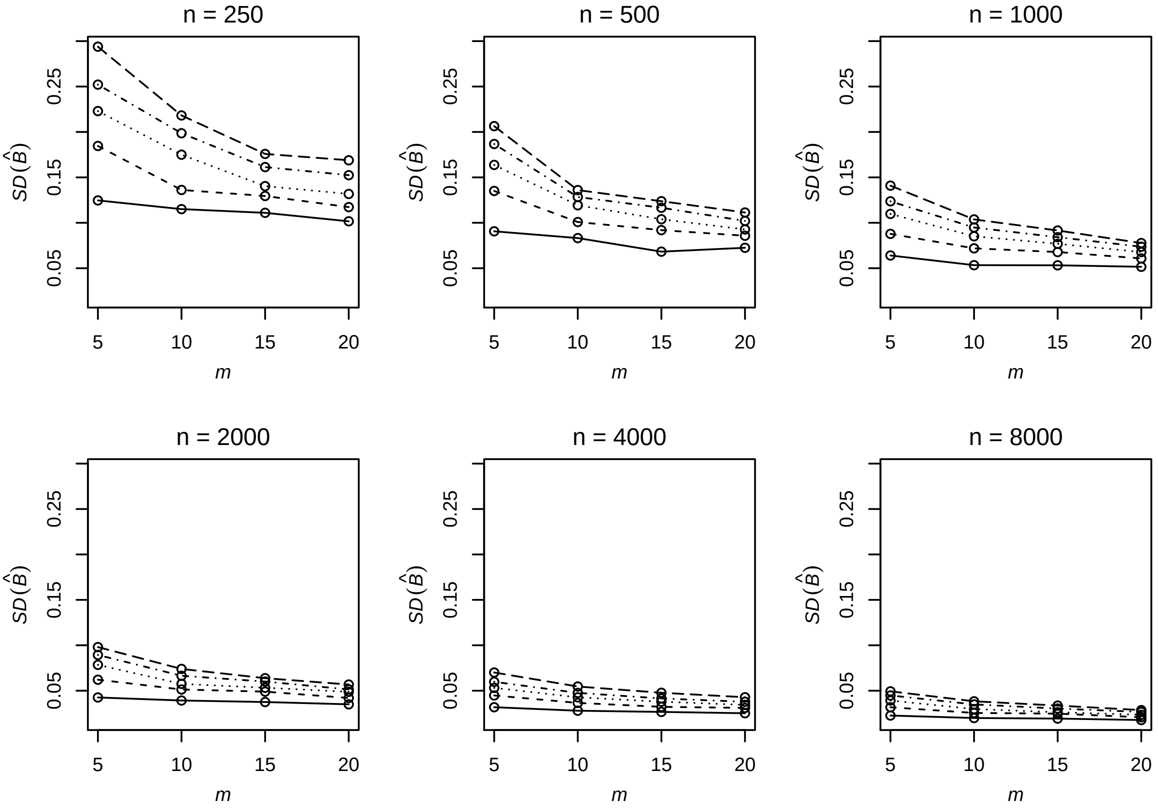


(a) Linkage plan 1.

(b) Linkage plan 2.

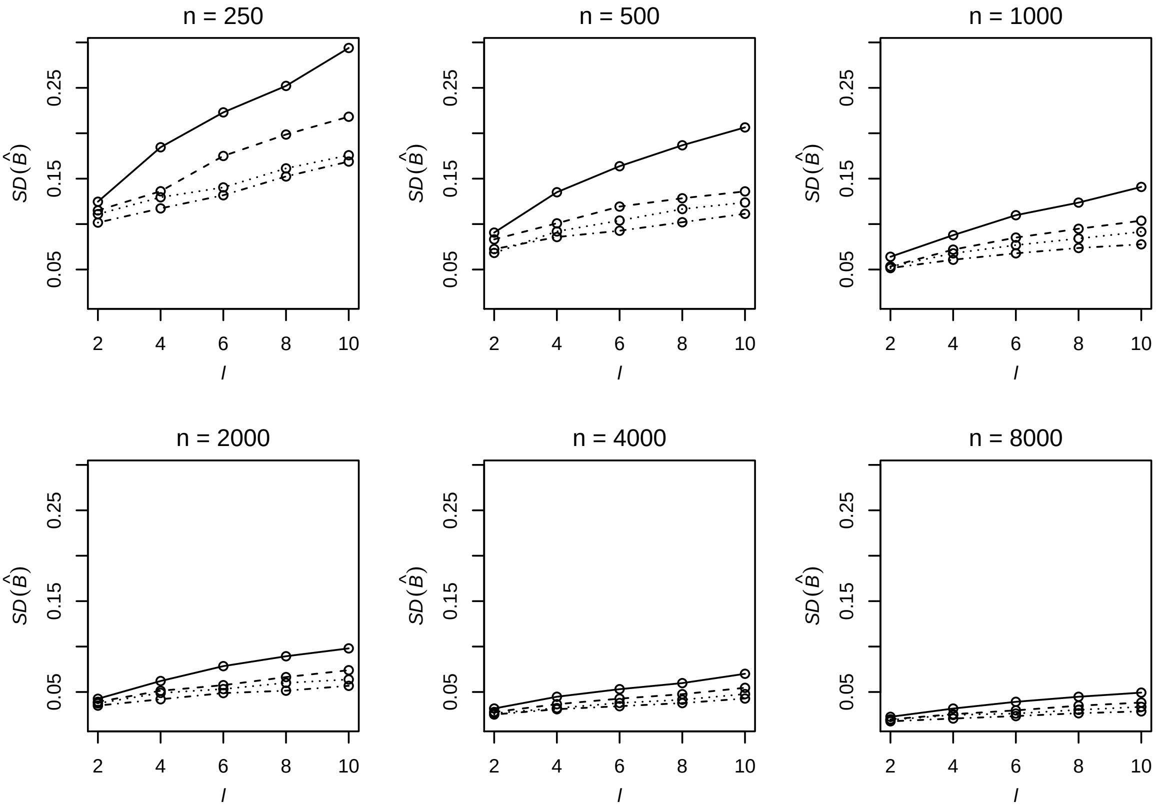
(c) Linkage plan 3.

Figure 2: Standard deviation of equating coefficient B against the number of common items for the Rasch model.



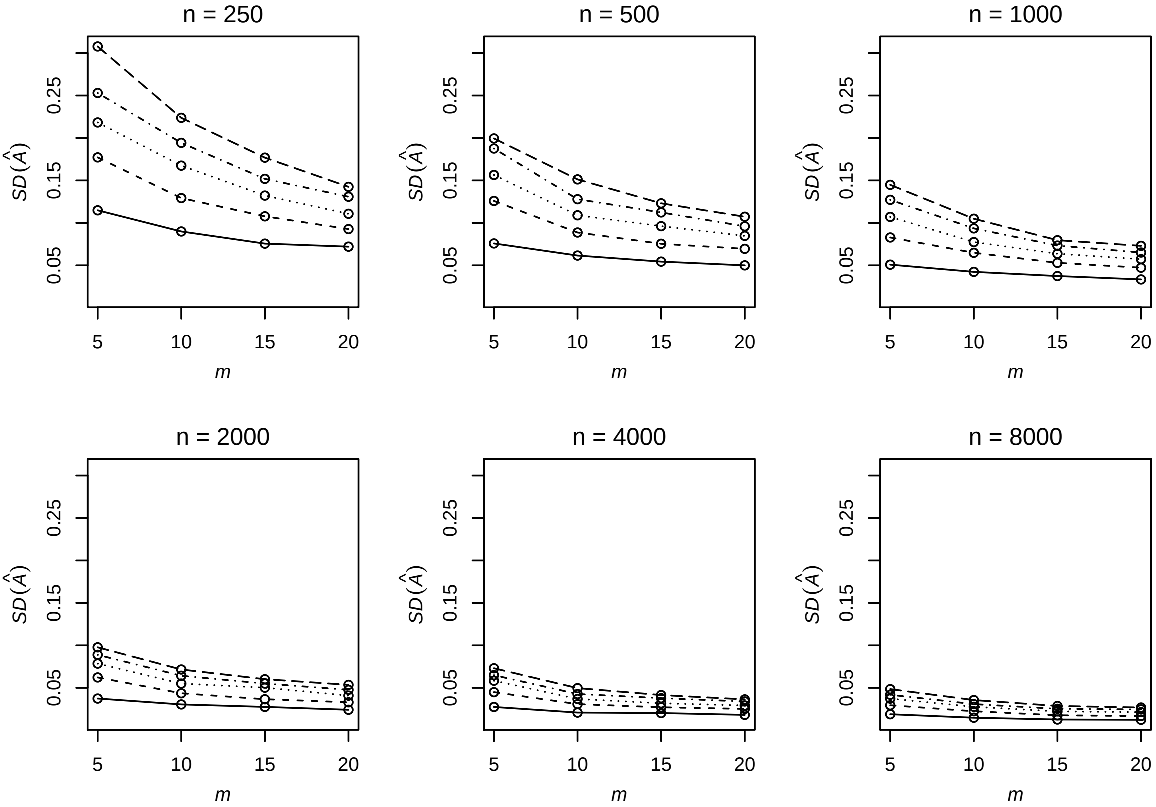
Lines refer to different lengths of the chain: solid line = 2, dashed line = 4, dotted line = 6, dotdash line =8, longdash line =10.

Figure 3: Standard deviation of equating coefficient B against the length of the chain for the Rasch model.



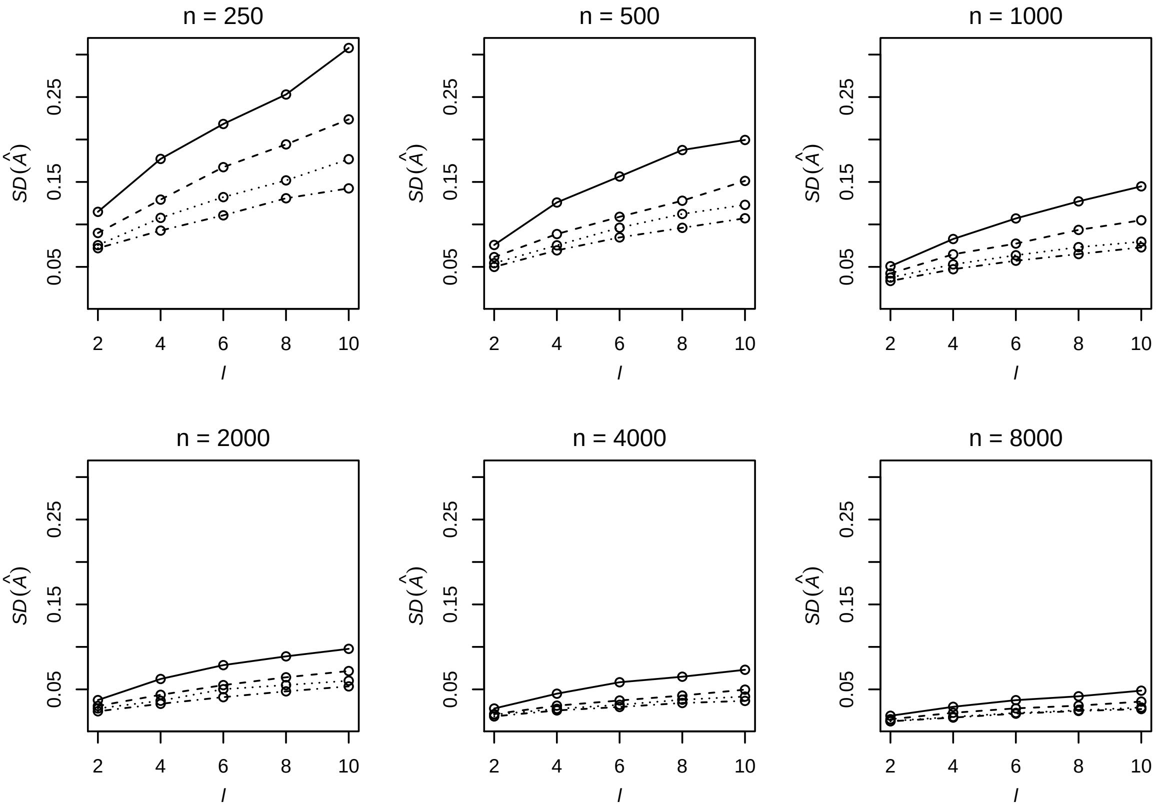
Lines refer to different numbers of common items: solid line = 5, dashed line = 10, dotted line = 15, dotdash line =20.

Figure 4: Standard deviation of equating coefficient A against the number of common items for the 2PL model.



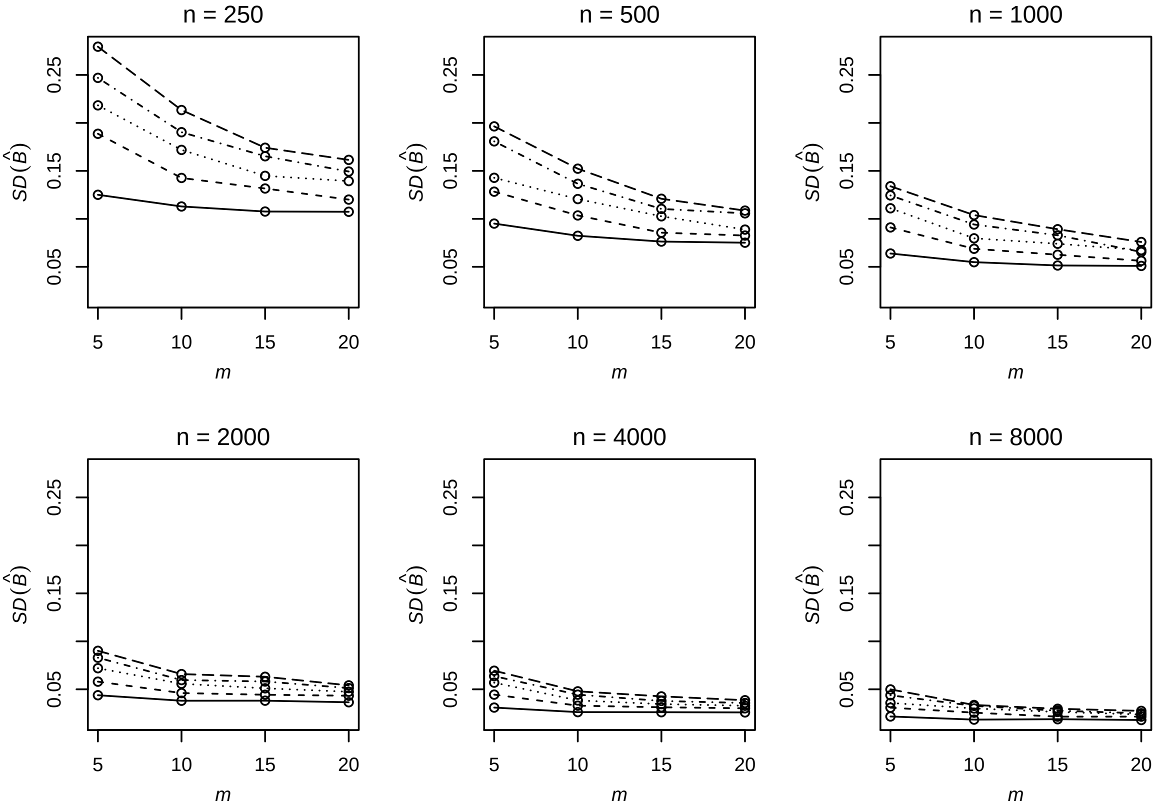
Lines refer to different lengths of the chain: solid line = 2, dashed line = 4, dotted line = 6, dotdash line =8, longdash line =10.

Figure 5: Standard deviation of equating coefficient A against the length of the chain for the 2PL model.



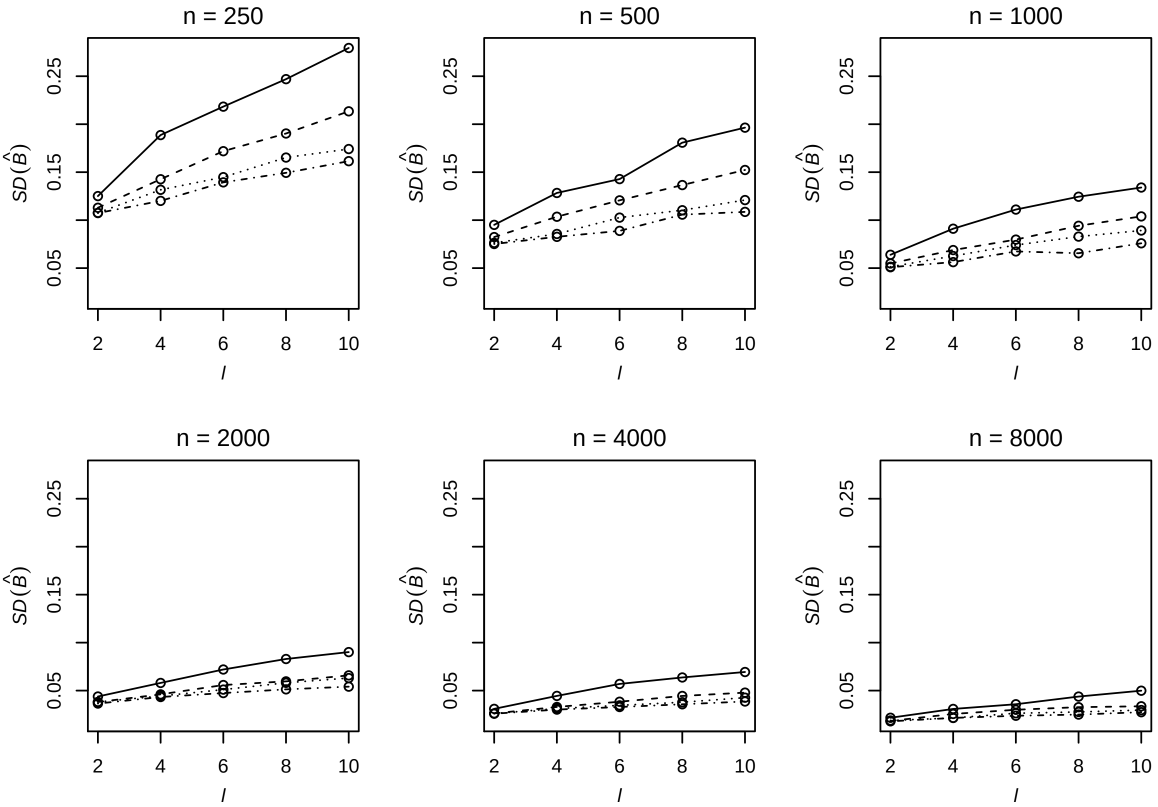
Lines refer to different numbers of common items: solid line = 5, dashed line = 10, dotted line = 15, dotdash line =20.

Figure 6: Standard deviation of equating coefficient B against the number of common items for the 2PL model.



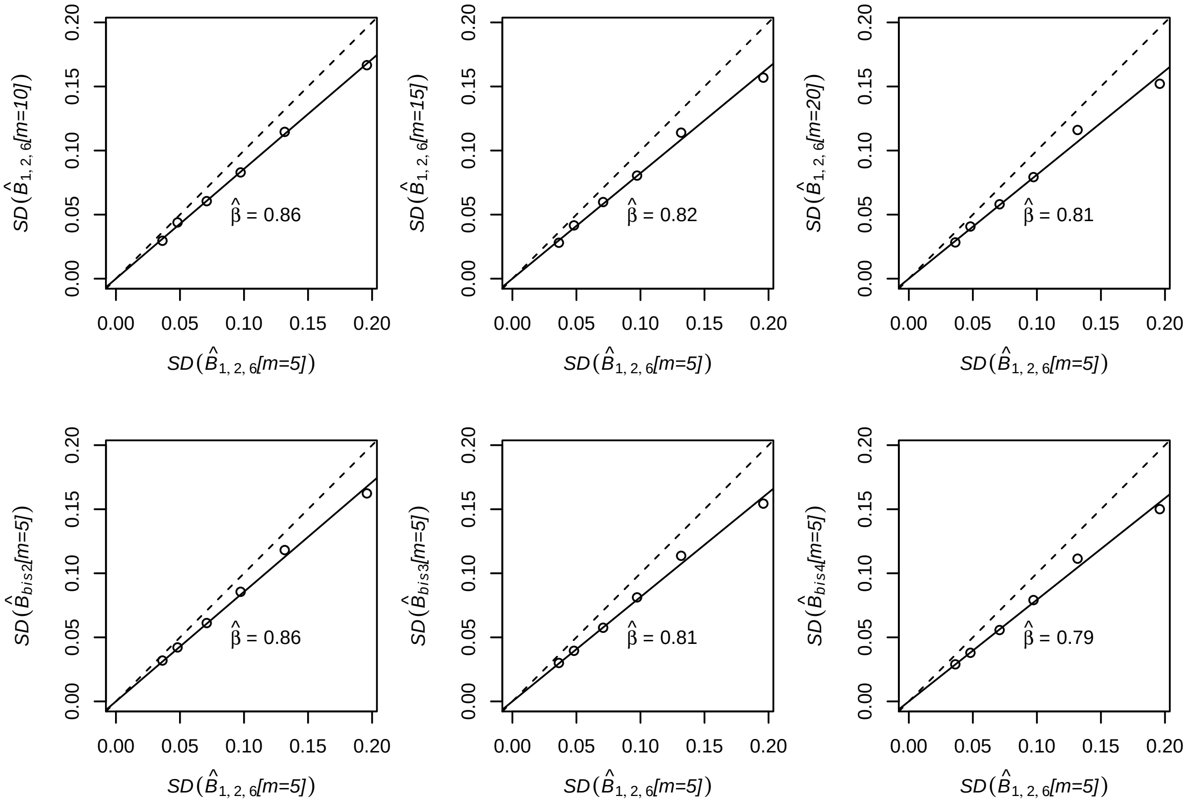
Lines refer to different lengths of the chain: solid line = 2, dashed line = 4, dotted line = 6, dotdash line =8, longdash line =10.

Figure 7: Standard deviation of equating coefficient B against the length of the chain for the 2PL model.



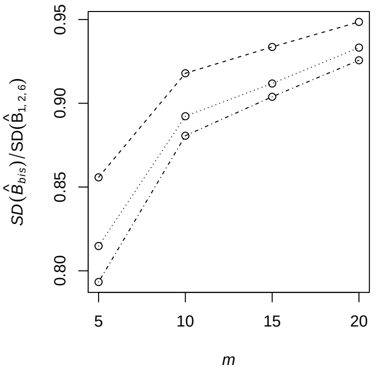
Lines refer to different numbers of common items: solid line = 5, dashed line = 10, dotted line = 15, dotdash line =20.

Figure 8: Comparison of the reduction of the standard deviation of the B equating coefficient for the Rasch model obtained by increasing the number of common items (on the top) and by averaging different chains with the bisector method (on the bottom).



The solid line represents the regression line with estimated coefficient $\hat{\beta}$.

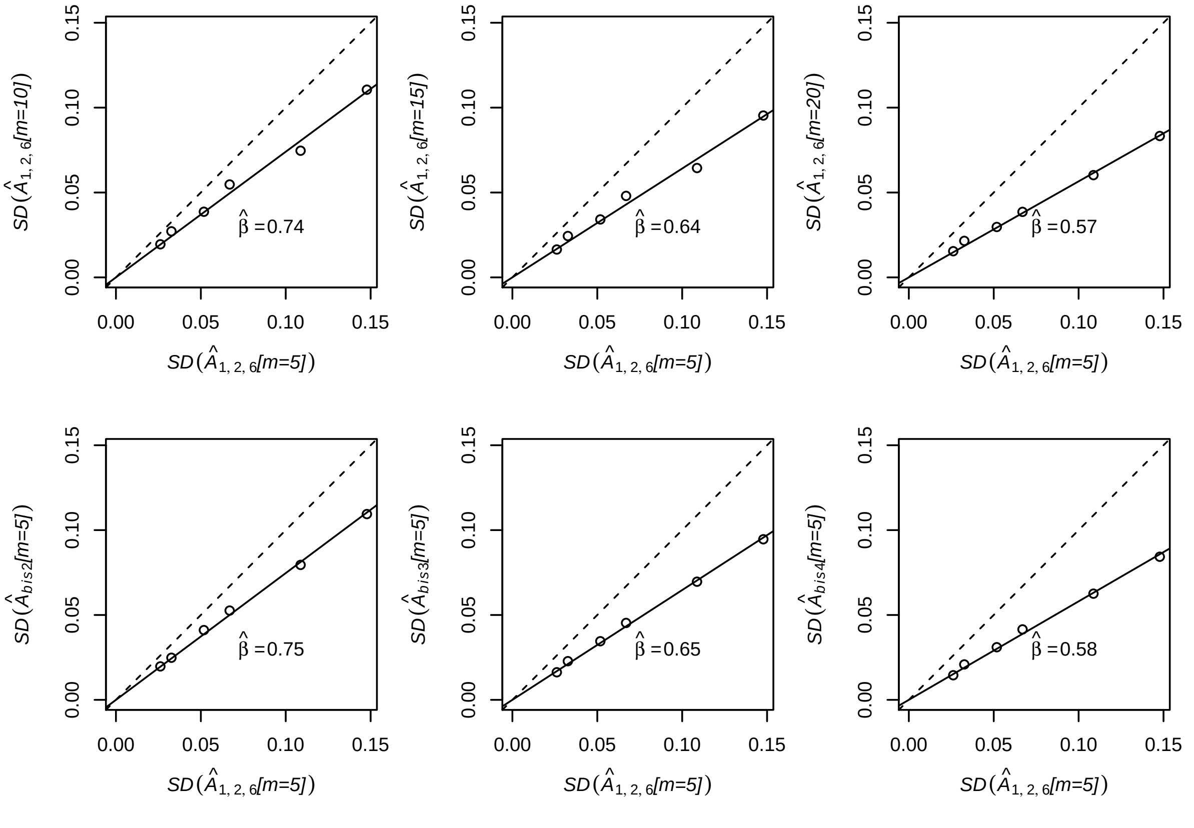
Figure 9: Estimated coefficients $\hat{\beta}$ representing the reduction of the standard deviation of the B equating coefficient for the Rasch model obtained by averaging different chains with the bisector method for the second linkage plan.



.

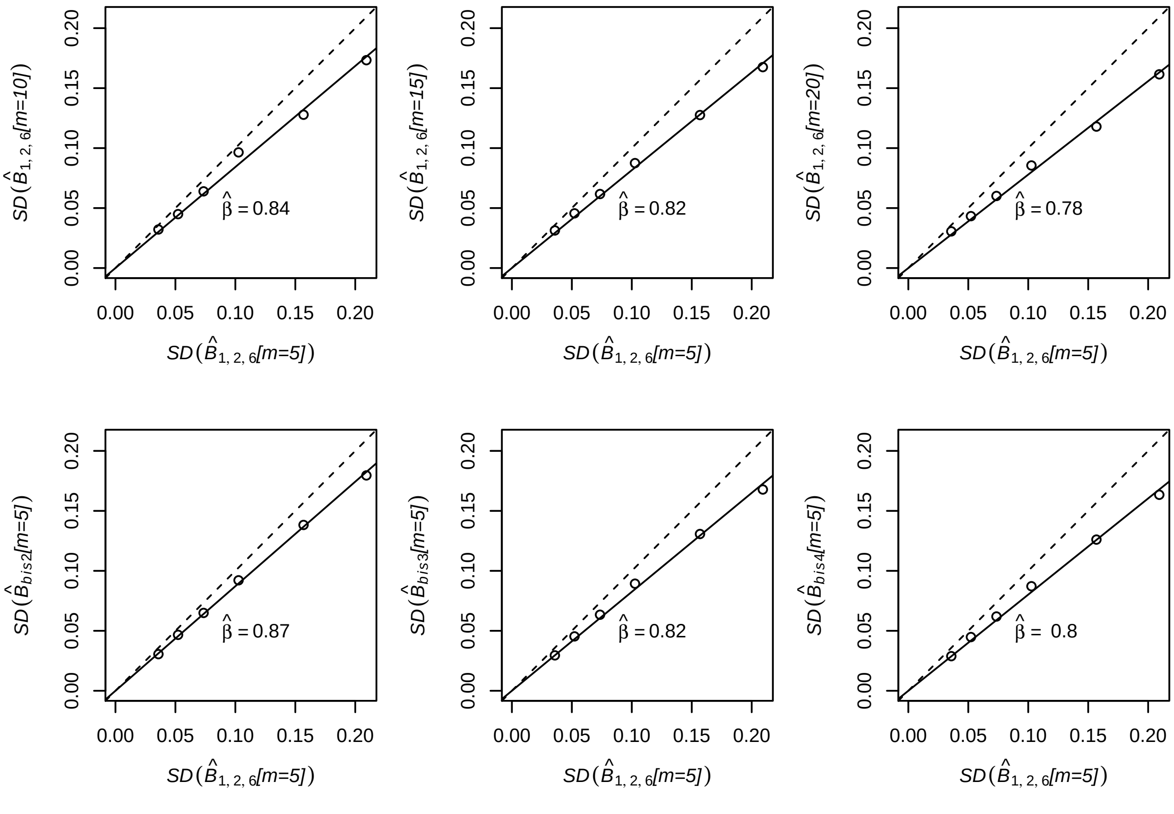
Lines refer to different number of chains averaged: dashed line = 2, dotted line = 3, dotdash line = 4.

Figure 10: Comparison of the reduction of the standard deviation of the A equating coefficient for the 2PL model obtained by increasing the number of common items (on the top) and by averaging different chains with the bisector method (on the bottom) for the second linkage plan.



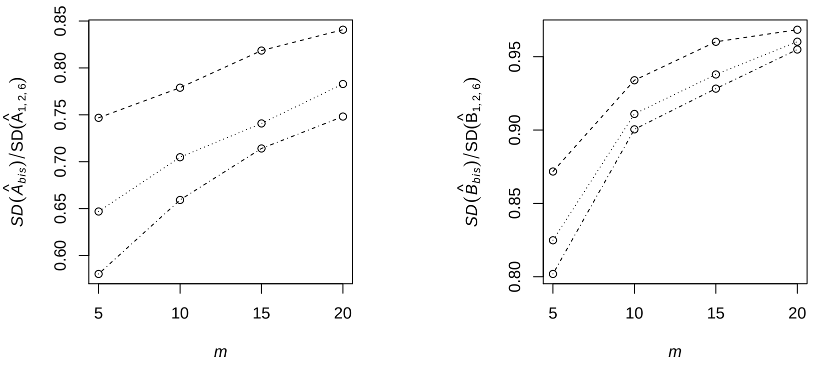
The solid line represents the regression line with estimated coefficient $\hat{\beta}$.

Figure 11: Comparison of the reduction of the standard deviation of the B equating coefficient for the 2PL model obtained by increasing the number of common items (on the top) and by averaging different chains with the bisector method (on the bottom) for the second linkage plan.



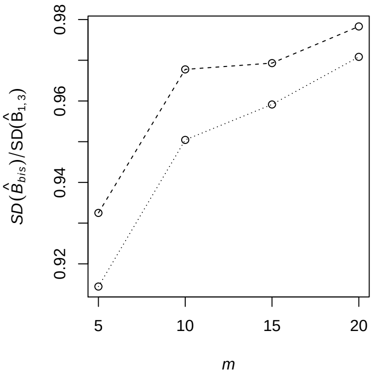
The solid line represents the regression line with estimated coefficient $\hat{\beta}$.

Figure 12: Estimated coefficients $\hat{\beta}$ representing the reduction of the standard deviation of the equating coefficients for the 2PL model obtained by averaging different chains with the bisector method for the second linkage plan.



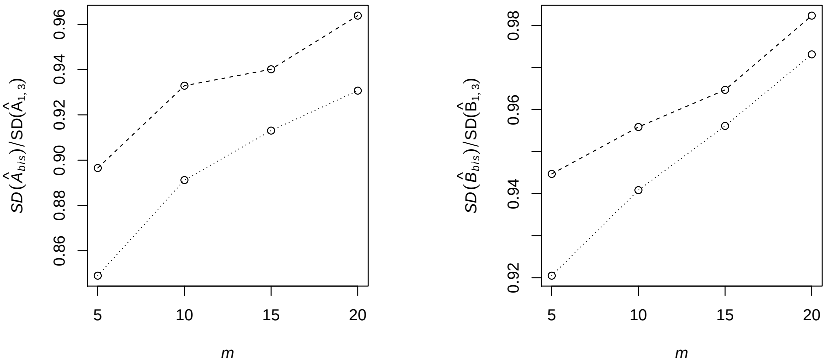
Lines refer to different number of chains averaged: dashed line = 2, dotted line = 3, dotdash line = 4.

Figure 13: Estimated coefficients $\hat{\beta}$ representing the reduction of the standard deviation of the B equating coefficient for the Rasch model obtained by averaging different chains with the bisector method in the third linkage plan.



Lines refer to the bisector method (dashed line) and the weighted bisector method (dotted line).

Figure 14: Estimated coefficients $\hat{\beta}$ representing the reduction of the standard deviation of the equating coefficients for the 2PL model obtained by averaging different chains with the bisector method in the third linkage plan.



Lines refer to the bisector method (dashed line) and the weighted bisector method (dotted line).

Table 1: Regression coefficients for the standard deviation of the B coefficient for the Rasch model.

n	250	500	1000	2000	4000	8000
Intercept	0.0938	0.0649	0.0490	0.0345	0.0235	0.0167
$1/m$	0.2059	0.1796	0.0903	0.0606	0.0510	0.0385
$l-2$	0.0039	0.0018	0.0018	0.0013	0.0015	0.0009
$1/m \cdot (l-2)$	0.0833	0.0606	0.0393	0.0282	0.0159	0.0125

Table 2: Regression coefficients for the standard deviation of the equating coefficients for the 2PL model.

(a) equating coefficient A

n	250	500	1000	2000	4000	8000
Intercept	0.0594	0.0420	0.0312	0.0204	0.0153	0.0103
$1/m$	0.3172	0.2255	0.1276	0.1165	0.0796	0.0535
$l-2$	0.0060	0.0050	0.0026	0.0026	0.0012	0.0013
$1/m \cdot (l-2)$	0.0884	0.0536	0.0459	0.0243	0.0219	0.0117

(b) equating coefficient B

n	250	500	1000	2000	4000	8000
Intercept	0.0975	0.0668	0.0442	0.0346	0.0231	0.0176
$1/m$	0.2024	0.1569	0.1276	0.0551	0.0519	0.0253
$l-2$	0.0036	0.0023	0.0021	0.0013	0.0005	0.0003
$1/m \cdot (l-2)$	0.0755	0.0538	0.0344	0.0231	0.0215	0.0156