Using the Distractor Categories of Multiple-Choice Items to Improve IRT Linking [*]

Jee–Seon Kim

University of Wisconsin, Madison

Paper presented at 2006 NCME Annual Meeting

San Francisco, CA

[*] Correspondence concerning this paper should be addressed to Jee-Seon Kim, Department of Educational Psychology, University of Wisconsin at Madison, 1025 Johnson Street, Madison, WI 53706. Electronic mail may be sent to jeeseonkim@wisc.edu.

Using the Distractor Categories of Multiple-Choice Items to Improve IRT Linking

Abstract

Simulation and real data studies are used to investigate the value of modeling multiple-choice distractors on item response theory linking. Using the characteristic curve linking procedure for Bock's (1972) nominal response model presented by Kim and Hanson (2002), all-category linking (i.e., a linking based on all category characteristic curves of the linking items) is compared against correct-only linking (i.e., linking based on the correct category characteristic curves only) using a common-item nonequivalent groups design. The correct-only linking is shown to represent an approximation to what occurs when using a traditional correct/incorrect item response model for linking. Results suggest the number of linking items needed to achieve an equivalent level of linking precision declines substantially when incorporating the distractor categories.

Key words: item response theory, linking, nominal response model

Using the Distractor Categories of Multiple-Choice Items to Improve IRT Linking

Recent research in item response theory (IRT) has highlighted the potential value of modeling the selection of distractor categories in multiple-choice items. Item response models such as the nominal response model (NRM; Bock, 1972) and the multiple-choice model (MCM; Thissen & Steinberg, 1984) provide tools by which the relationship between ability and the selection of specific response categories in multiple-choice items can be studied. As most multiple-choice items are written so as to contain one correct response, these so-called "choice" IRT models thus provide information about the relative attractiveness of the remaining distractor categories to examinees of different ability levels. The value of attending to distractor selection has now been illustrated for several educational measurement applications, including for example, ability estimation (Baker & Kim, 2004), interpretation of differential item functioning (Thissen, Steinberg, & Wainer, 1993), and cognitive diagnosis of item response patterns (Bolt, Cohen, & Wollack, 2001).

Another potential, but yet unexplored, advantage of choice IRT models is the greater information they provide for purposes of IRT linking. In this paper, IRT linking refers to the process by which estimates from separate IRT analyses are put on a common scale. Linking is of critical importance in IRT, as multi-sample analyses are commonplace in educational measurement. Ultimately, linking is the process by which the hallmark feature of item response theory, namely model parameter invariance, is put to practical use. Accurate linking is thus necessary for a wide range of testing applications, including test score equating, differential item functioning, and computerized adaptive testing.

Various data collection designs permit some form of IRT linking (see Kolen and Brennan, 2004). The current study focuses on a design involving two examinee groups (possibly with different ability distributions) administered test forms with common items, often referred to as a *common-item non-equivalent groups* design. Because the common items are the basis for IRT linking, an important consideration is the number of such items needed to ensure a linking with adequate precision. In this paper, we refer to the common items used for IRT linking as *linking items*. From a purely statistical standpoint, the larger the number of linking items, the more precise the linking. Previous simulation studies of this issue have demonstrated consistent improvements in linking precision due to increasing the number of linking items up to fifty (Kim & Cohen, 2002), although other studies appear to suggest minimal gains beyond ten linking items (e.g., Hanson & Béguin, 2002). Naturally the value of increasing the number of linking items on linking precision might be expected to be affected by factors such as sample size, the item parameters associated with the linking items, and/or the examinee ability distributions. Moreover, other characteristics of the linking items are also important to consider in actual applications. For example, in differential item functioning (DIF) analyses, it is critical to use as linking items only items known to lack DIF (Millsap & Everson, 1993), while in equating applications, the content representativeness of the linking items is important to consider (Kolen & Brennan, 2004).

Linking methods for polytomous IRT models, such as the graded response model (Baker, 1992), are often straightforward extensions of methods originally developed for dichotomous IRT models, such as the 2PL. Recently, Kim and Hanson (2002) presented a characteristic curve linking procedure based on the multiple-choice model (MCM; Thissen & Steinberg, 1984) that included some corrections to previous formulae (Baker, 1993) for characteristic curve linking under the nominal response model (NRM). Kim and Hanson's procedure can be viewed as a generalization of the characteristic curve linking procedure of Haebara (1980) for IRT calibrations based on either the MCM or NRM. The approach uses all response categories (including distractors) in determining the linking transformation, and for this reason may provide greater accuracy than linking methods based on only the dichotomous responses.

Despite the anticipated benefits of incorporating distractor categories into IRT linking, the ultimate effects are not entirely clear. Indeed, there may be reasons to believe that attending to such information could be negligible, or even detrimental to linking. First, because each distractor option is typically selected by a much smaller number of examinees than the correct response, the parameters related to distractor categories are frequently estimated with much less accuracy than the parameters for the correct response option. Second, the distractors generally provide most of their information at lower ability levels, making their contributions to the overall linking of IRT calibrations less obvious. Third, because the estimation of distractor category parameters is dependent to a large extent on the estimation of the other category parameters (including the correct category), the distractors should not provide information independent of that already provided by the correct response.

The goal of this study is to examine the degree to which the NRM may improve IRT linking through its incorporation of all response categories into the linking process. Specifically, we examine the correspondence between the numbers of linking items that must be used to achieve an equivalent level of linking precision when linking based on all response categories versus only attending to the correct category.

IRT Linking Under the Nominal Response Model

Bock (1972) introduced the NRM as an item response model for multiple choice options. Under the NRM, the probability of selecting a response category k is modeled as a function of a unidimensional ability level θ:

$$P(X_{ij} = k|\theta_i; a_{jk}, b_{jk}) = \frac{\exp(a_{jk}\theta_i + b_{jk})}{\sum_{h=1}^{K} \exp(a_{jh}\theta_i + b_{jh})},$$

where X_{ij} represents the response choice of examinee i on item j, a_{jk} and b_{jk} the slope and intercept parameters, respectively, for category k on item j, and θ_i the latent ability level of examinee i. The a_{jk}s account for the relationship between ability and propensity to select category k, while the b_{jk}s control the overall propensity to select category k

at $\theta = 0$, often the mean of the ability distribution. For identification purposes, the constraints $\sum_{k=1}^{m_j} a_{jk} = 0$ and $\sum_{k=1}^{m_j} b_{jk} = 0$, are frequently imposed across categories within an item, and will also be used in the current study.

Compared to a dichotomous IRT model, such as the 2PL, which models only the correctness of the item response, the NRM models all response options. Each response category can be associated with a curve, called a category characteristic curve, that represents the probability of selecting the category as a function of θ. Figure 1 provides an illustration of NRM category characteristic curves for a five-choice item, as well as the item characteristic curve for a corresponding 2PL item. For each NRM item, there always exists one response category that strictly increases as a function of θ (generally the correct response category), and one that strictly decreases. For many multiple-choice items, the existence of the latter type of category may not be expected; models such as the MCM of Thissen and Steinberg (1984) may be more realistic for such items. However, Drasgow, Levine, Tsien, Williams and Mead (1995) have found the NRM to provide a good fit to actual data from a multiple-choice test.

As for dichotomous IRT models, any NRM solution is identified up to a linear transformation of the ability scale. Consequently, when two tests measuring the same ability are estimated separately, and the model fits both tests well, the resulting solutions will be related by a linear transformation. Assuming we wish to transform the solution for an equating form to that of a base form, we can determine the linear transformation when there exist common items or examinees across forms. We will refer to the IRT estimates for the base form as the target solution, and the underlying ability metric as the target metric for linking. Then the corresponding abilities for the target solution should be related to the abilities for the equating solution as : $\theta^* = S\theta + I$, where θ^* denotes the ability for an equating population examinee once place on the target metric, and S and I denote the linking slope and linking intercept parameters, respectively. Transformations of the item category slope and intercept parameter estimates for the equating solution are based on the same linking slope and intercept parameters. Specifically,

$$\hat{a}^*_{jk} = \frac{\hat{a}'_{jk}}{S}$$

and

$$\hat{b}^*_{jk} = \hat{b}'_{jk} - \frac{\hat{a}'_{jk}}{S} I = \hat{b}'_{jk} - \hat{a}^*_{jk} I,$$

where $\hat{a}'_{jk}$ and $\hat{b}'_{jk}$ represent the category slope and intercept estimates of the equating solution on their original metric, and $\hat{a}^*_{jk}$ and $\hat{b}^*_{jk}$ represent the category slope and intercept estimates of the equating solution on the target metric. Accurate estimation of the linking parameters is thus critical not only for ensuring comparability of the ability levels of equating form and base form examinees, but also the characteristics of equating and base form items. Failure to adequately recover these parameters naturally leads to bias in comparisons of examinees and/or items administered with different test forms.

Various methods exist for estimating the linking parameters I and S. The *characteristic curve method* estimates the linking parameters by minimizing the sum of the squared differences of the item category characteristic curves for the linking items from the target and equating solutions. Specifically, the method minimizes

$$F = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [P(\theta_i | \hat{a}_{jk}, \hat{b}_{jk}) - P(\theta_i | \hat{a}^*_{jk}, \hat{b}^*_{jk})]^2, \tag{1}$$

where $\hat{a}_{jk}$ and $\hat{b}_{jk}$ denote the corresponding parameter estimates of linking items from the equating solution, M is the total number of response categories accumulated across items, and i indexes several locations along the θ scale. As shown in Kim and Hanson (2002), estimates of the linking intercept parameter I and the linking slope parameter S can be determined by solving a system of equations in which the derivatives of F with respect to I and S are both set to zero (Dennis & Schnabel, 1996). These derivatives are given by:

$$\frac{dF}{dS} = -\frac{2}{NM} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{m_j} [P(\theta_i | \hat{a}_{jk}, \hat{b}_{jk}) - P(\theta_i | \hat{a}^*_{jk}, \hat{b}^*_{jk})] \left(\frac{d}{dS} P(\theta_i | \hat{a}^*_{jk}, \hat{b}^*_{jk}) \right)$$

and

$$\frac{dF}{dI} = -\frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{k=1}^{m_j}[P(\theta_i|\hat{a}_{jk},\hat{b}_{jk}) - P(\theta_i|\hat{a}^*_{jk},\hat{b}^*_{jk})]\left(\frac{d}{dI}P(\theta_i|\hat{a}^*_{jk},\hat{b}^*_{jk})\right),$$

which include corrections to the derivatives originally specified by Baker (1993). Additional details on the minimization procedure are described in Kim and Hanson (2002), who also report on simulation and real data analyses that support use of the method with the NRM.

Linking Based on Only the Correct Categories

Multiple-choice item response data are often analyzed using dichotomous item response models applied to correct/incorrect item responses. IRT linking procedures, including the characteristic curve method, can be applied using models such as the two-parameter logistic (2PL) or three-parameter logistic (3PL) models. Relative to the NRM, we might view such models as only attending to the correct response categories when minimizing the function F. However, it is also important to note that models such as the 2PL are statistically different and technically incompatible with the representation of the correct response probability modeled by the NRM. That is, the item category curve associated with the correct response category in the NRM will not be identical to the curve obtained when fitting a model such as the 2PL, as the NRM does not possess a score collapsibility property. Nevertheless, the approximation is generally quite close. As a result, we might expect a fairly close approximation to the performance of 2PL linking by performing an NRM linking based only on the correct response category, a result we verify shortly.

Following Kim and Hanson (2002), a characteristic curve linking based only on the correct response categories of the NRM can be determined by minimizing the loss function:

$$F = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}[P(\theta_i|\hat{a}_{j,k(j)},\hat{b}_{j,k(j)}) - P(\theta_i|\hat{a}^*_{j,k(j)},\hat{b}^*_{j,k(j)})]^2, \tag{2}$$

where $k(j)$ now denotes the correct response category for item j, and all other terms are as in Equation (1).

The first partial derivatives of F with respect to the two linking parameters I and S result in

$$\frac{dF}{dS} = -\frac{2}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}[P(\theta_i|\hat{a}_{jk(j)}, \hat{b}_{jk(j)}) - P(\theta_i|\hat{a}^*_{jk(j)}, \hat{b}^*_{jk(j)})]\left(\frac{d}{dS}P(\theta_i|\hat{a}^*_{jk(j)}, \hat{b}^*_{jk(j)})\right)$$

and

$$\frac{dF}{dI} = -\frac{2}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}[P(\theta_i|\hat{a}_{jk(j)}, \hat{b}_{jk(j)}) - P(\theta_i|\hat{a}^*_{jk(j)}, \hat{b}^*_{jk(j)})]\left(\frac{d}{dI}P(\theta_i|\hat{a}^*_{jk(j)}, \hat{b}^*_{jk(j)})\right).$$

The same minimization procedure used by Kim and Hanson (2002) can be applied with respect to Equation (2). It should be acknowledged that this linking strategy is not likely to be one actually used in practice; its value in the current study comes from the expectation that this type of linking should closely approximate linking based on dichotomous IRT models, while also being based on the same model used for the usual NRM based-linking. Differences between the estimates of I and S obtained using Equation (1) versus Equation (2) can thus be directly attributed to the role of the distractor options in the NRM linking. In the context of a simulation study, it becomes possible to evaluate whether these differences imply better or worse recovery under the NRM, as well as to quantify the difference between the two methods. Because the linking following the current procedure is based only on the correct response categories, we refer to it as ‘correct-only’ (CO) linking. By contrast, a linking based on all response categories of the linking items will be referred to as ‘all-category’ (AC) linking.

Simulation Study

The simulation study considered a thirty-six item test with NRM parameters reported in Table 1. The parameters are based on NRM estimates from a 36-item college level mathematics placement test administered to 15,386 examinees (Center for Placement Testing, 1998). The test is taken by entering first-year students in the university of Wisconsin system to assist in course placement decisions, and includes items that represent a wide range of content areas, including high school algebra, geometry, and trigonometry. Each item contained five response categories, exactly one of which represented a correct response. The estimates reported in Table 1 were obtained using the software package MULTILOG 7 (Thissen, Chen, & Bock, 2003).

The goal of the simulation study was to evaluate the degree to which recovery of the linking parameters (I and S) was improved when using AC versus CO linking based on the NRM. In order to quantify the difference between methods, both AC and CO linkings were studied using varying numbers of linking items, ranging from one to twenty in increments of one. It was anticipated that for a given number of linking items, the effects of AC linking could be determined by examining the corresponding number of linking items under CO linking needed to return the same level of linking precision.

In addition to the number of linking items, the simulation varied two additional factors: sample size of the equating sample and the ability distribution of the equating population. Sample size for the equating sample was considered at levels of 250, 500, 1,000, and 3,000. Fifteen different ability distributions were considered for the equating population. In each case, a normal distribution of ability was assumed. In all linkings, the NRM solution for the target population was set at the values reported in Table 1, and could therefore be regarded as a population with ability mean of zero and variance of one. By default, the ability mean and variance for each MULTILOG run are also set at 0 and 1 for the equating sample calibrations. Consequently, the generating parameters of the equating population ability distribution also determine the "true" linking parameters needed to put the equating sample solution on the metric of the target population, as will be explained in more detail below. The mean and standard deviation of ability for the equating populations were considered at levels of -1, -0.25, 0, 0.25, and 1; and 0.5, 1, and 2, respectively.

For each combination of sample size and equating calibration ability distribution conditions (a total of 4 × 15 = 60), a total of 1,000 datasets were simulated from the NRM. Each such dataset was then estimated using the NRM, resulting in NRM estimates for 36 items.

Each of the 1,000 datasets could then be linked to the target solution using some subset of the thirty-six items as linking items. However, so as to avoid confounding results for the number of linking items with effects due to the parameter values of the specific items chosen for the linking, a different subset of linking items was chosen for each of the 1,000 datasets. The subset of linking items chosen was determined randomly. For example, in the two linking-item condition, the linking items chosen from the first dataset might be items 12 and 27, while for the second dataset it would be 5 and 18, and so on.

Simulation Study Results

Table 2 reports the linking parameter recovery results for the I and S linking parameters under the equating sample size=1,000 condition under the different equating population ability distribution conditions. Similar patterns of results across conditions were observed for the other sample size levels. It should be noted that the true I and S linking parameters correspond exactly to the mean and standard deviation, respectively, of the equating population.

Recovery results are reported separately for several different numbers of linking items (2, 5, 10, 15, 20). The root mean square error (RMSE) reported in each cell represents the square root of the average (across the 1,000 simulated linkings) squared difference between the estimated linking parameter and the true linking parameter. Note that because the mean and variance of the equating sample are not exactly the mean and variance of the equating population, the true linking parameters used in computing the RMSE varied slightly across the 1,000 datasets. In all cases, the true linking parameters used in computing the RMSEs were based on the sample means and standard deviations, respectively, so as not to bias the results.

As expected, linking parameter recovery improves as the number of linking items is increased, but appears to asymptote at about ten items. Linking parameter recovery is also affected by the equating population ability distribution parameters, with poorer recovery occurring when the equating population has less variance (S=0.5).

Most important for the current analysis, however, is the comparison between AC and CO linkings. Across all equating populations, there appears to be a consistent benefit to the AC linking. The differences between the AC and CO linkings are most dramatic when the number of linking items is smallest (i.e., 2), in which case the RMSE of the intercept under AC linking is on average approximately 27% that of the CO linking, while for the slope it is 37%. AC linking remains superior even as the number of linking items becomes large (20), although the difference from CO linking becomes noticeably smaller, and perhaps negligible, when number of linking items exceeds 10. Somewhat surprisingly, the proportional benefit in the AC versus CO linking appears to remain relatively constant across different equating population ability distributions. That is, even when the equating population is of high ability (and where a smaller overall proportion of examinees will be selecting distractors), there remains a clear benefit to the AC linking, although the recovery under each of the AC and CO linkings appears worse than for other equating populations.

When the effect of AC linking is evaluated in terms of number of linking items, it would appear that an AC linking based on five linking items produces results comparable to a CO linking using twenty linking items. Interestingly, across all equating population conditions, the AC linking with five linking items produces almost identical results to that for the CO linking with twenty linking items in terms of recovery for both I and S.

Figures 2a and 2b illustrate the RMSEs of the linking parameters (now averaged across the fifteen equating population conditions) for each of the sample size levels. In these figures, the results are plotted as a function of all levels of number of linking items, which ranged from one to twenty in increments of one. From these figures it can be seen that as sample size increases, linking recovery naturally improves, as should be expected given the better estimation of the item parameters. More interesting, however, is the clear superiority of the AC linking even with only one linking item. Specifically, one linking item under AC appears to be roughly equivalent to use of four linking items when linking under CO. Consequently, the 1:4 ratio observed in Table 2 appears to hold roughly across different numbers of linking items. That is, under AC linking it appears that nearly equivalent results emerge as for the CO linkings when using four times as many linking items.

As noted earlier, the CO linking is not presented here as a practical linking method when using the NRM, but rather as an approximation to linking precision as would occur when using a dichotomous model such as the 2PL. In order to verify that the precision of CO linking provides a close approximation to what occurs when linking is based on a model such as the 2PL, the simulation conditions considered here were replicated using the 2PL model. The item parameters for data simulation were based on 2PL estimates from the same math placement data that were the basis for the NRM estimates in Table 1. Linking parameter recovery results were considered for the 2PL using 2, 5, 10, 15, and 20 linking items, the same conditions displayed in Table 2. A comparison of results in terms of linking parameter recovery for the CO linking and 2PL linking are shown in Table 3. The nearly identical findings support the earlier claim that the CO linking appears to closely approximate what occurs when linking using dichotomous IRT models.

It might be argued that part of the success of the AC linkings above can be attributed to the fact that the NRM was used to generate the data. Hence, the value of distractor options in linking may be overstated due to their perfect fit by the NRM. The next study addresses this limitation by comparing AC and CO linkings of the NRM when applied to real test datasets.

Real Data Study

In this study, actual data were used from the mathematics placement test that was the basis for the estimates reported in Table 1. A random sample of 3,000 examinees was selected from the full dataset of 15,123 examinees to provide an NRM solution that would function as a target solution for all of the analyses. From the remaining 12,123 examinees, additional samples of up to 3,000 examinees were selected as equating samples. Nine different equating populations were specified in terms of their ability distributions. To create an equating sample from a particular ability distribution, the following procedure was followed. First, a total correct score was determined for each examinee in the full dataset. These total scores were then standardized. An ability distribution for an equating sample was then specified in terms of the distribution of standardized scores. In all cases these distributions were normal, but varied in terms of their mean and variance. Nine different ability distributions were considered for the equating samples by crossing standardized test score means of -1, 0 and 1 with variances of 0.5, 1, and 2. Next, for each of the nine specified equating samples, each of the 12,123 examinees could be assigned a likelihood of selection based on the normal density evaluated at the examinee's standardized test score. Finally, examinees for each of the equating samples were randomly selected with probabilities proportional to these likelihoods. An initial sample of 3,000 was extracted for each of the nine equating distributions from the 12,123 examinees. (It should be noted that although each sample was selected without replacement, there is some overlap of examinees across the nine equating distribution samples.) From each of the nine equating distribution samples of 3,000, sample size conditions of 250, 500 and 1,000 were also considered by sampling from the 3,000 examinee dataset. Specifically, a random sample of 1,000 was chosen from the sample of 3,000, a random sample of 500 from the sample of 1,000, and so on.

All thirty-six items used to obtain the target solution and equating solutions were common because they were based on the same form. To simulate a realistic linking situation, however, only a subset of the thirty-six were used as linking items. The remaining common items could then be used to evaluate the success of the linking. Specifically, we considered the root mean square difference (RMSD) of the target solution and equating solution item parameter estimates once the solutions had been linked.

This alternative criterion for evaluating the success of the linking was necessary because, unlike the simulation study, the true linking parameters are unknown. Nevertheless, due to the invariance properties of the NRM, we should suspect that an accurate linking will make these estimates nearly identical.

For each of the 9 (equating population) $\times$ 4 (sample size) = 36 datasets, an NRM solution was obtained using MULTILOG, again with ability mean of 0 and variance of 1. As in the simulation study, linkings were performed using both AC and CO linkings, with the number of linking items ranging from 1 to 20.

For each of the 36 equating solutions, a total of 100 linkings were performed for each number of linking items condition. As in the simulation, each linking involved a random selection of items from among the 36 to serve as linking items. The remaining items were then used to evaluate the accuracy of the linking. Average RMSDs were computed across the 100 linkings and for all categories of common items not used for the linking. Separate averages were computed for the item category intercepts and item category slopes.

Real Data Study Results

Table 4 reports the RMSD between parameter estimates for the common non-linking items under the equating sample size=1,000 condition. Unlike the simulation study, these values do not appear to go to zero even as the number of linking items increases, but instead would appear to asymptote at some value above zero. This can be attributed to at least a couple of factors. First, the RMSDs are a function of two sets of parameter estimates (as opposed to the comparison of estimates against true parameters in the simulation study), and thus for fixed sample sizes should be more affected by estimation error. Second, to the degree that the NRM fails to fit the data, we may expect some lack of item parameter invariance, as different ability distributions will often lead to different item parameter estimates when an IRT model fails to fit (see e.g., Bolt, 2002). As a result, even as sample sizes increase, we expect some differences to remain even among item parameter estimates successfully linked.

Nevertheless, across both the different equating sample conditions and number of linking item conditions, it again appears that the AC linking consistently outperforms the CO linking, with lower RMSDs observed for both the item category slopes and item category intercepts across all conditions. As in the simulation, the difference also appears to diminish as the number of linking items increases. Although the percentage reduction in RMSD for AC versus CO appears lower than for the RMSEs in the simulation study, this smaller difference can be attributed to a couple of factors. First, recovery in the current analysis is being evaluated with respect to item parameter estimates, whereas linking parameter estimates were considered in the simulation. In general, the item parameter estimates will be more affected by sampling error. Second, for the reasons stated above, the use of real data lowers the reachable upper bound in terms of estimation accuracy.

For these reasons, a better way to compare the AC and CO linkings would again be to compare their relative accuracy in relation to the number of linking items. As in the simulation, it appears that AC linkings achieve levels of precision that require a much larger number of linking items under CO linkings. In this case, the results for AC linkings based on five linking items produce nearly equivalent results to those obtained for CO linkings based on 15 items, at which point the CO linking appears to reach an asymptote.

As had been observed in Table 2 for the simulation, the relationship in linking precision between AC and CO linking (approximately 1:3 to 1:4 in terms of linking items) generally holds quite well across the different equating population conditions, including those involving equating samples of higher ability.

Figures 3a and 3b illustrate the RMSDs between target and equating solutions across the different sample size conditions, now averaged for the nine different equating populations. For both AC and CO linkings, it can again be seen that the results (even under sample sizes of 3,000) asymptote at a level above zero. In comparing the AC and CO linkings, it again appears that AC is usually better, although for situations involving one linking item, there are a couple of conditions where the AC linking may have been slightly worse. Nevertheless, the results appear quite encouraging for considering distractor categories in the process of linking.

Discussion and Conclusion

Item response models used for educational and psychological measures vary considerably in their complexity. For multiple-choice items specifically, practitioners often model only the correctness of the response using a dichotomous outcome IRT model (e.g., the 2PL or 3PL). A review of studies applying IRT to multiple-choice items suggests that linkings based on the dichotomous responses remain more popular than methods that incorporate the distractor categories.

However, due to the dependence of many IRT applications on accurate IRT linking, this paper suggests that fitting more complex models such as the NRM, assuming they provide a reasonable fit to the data, can be of considerable benefit. Both the simulation and real data analyses are quite promising in support of the use of distractor categories for reducing the number of linking items needed to achieve good linking precision.

Based on the current analyses, the value of incorporating distractor categories into linking appears greatest when the available number of linking items is small. The simulation study, in particular, suggests a very substantial increase in linking accuracy, with each additional distractor option providing nearly the equivalent of an additional linking item when linking only on the correct responses. Future study might examinee the degree to which these results can be replicated using other models for distractor options, such as the MCM.

It is important to mention, however, that this benefit becomes substantially reduced when the available number of linking items is large (10 or greater). Because such designs are common in many testing programs, it might be questioned whether the slight improvement in linking precision is actually of practical benefit. It may well be that for these conditions, the cost of fitting a more complex model such as the NRM outweighs the modest gains in linking precision.

Somewhat surprisingly, where the benefit exists, the usefulness of distractors for linking appears to be largely unaffected by the ability distribution of the equating population. Even in high ability populations, where distractor selection is less common and where estimation of parameters related to distractor selection is less accurate, there appears to be information in distractor selection that can assist in linking calibrations.

There are several limitations to the current study. First, as noted earlier, the simulation results assume a perfect fit of the NRM to the data, and thus potentially overstates the value of the NRM when applied to real data. As suggested by the real data study, the value of modeling all response categories may be diminished somewhat when taking into account the occurrence of model misfit. In particular, the performance of the NRM when linking with respect to only one linking item, would appear to be questionable (although for other reasons, one-item linking would not be advisable).

Second, the ability distributions considered in the simulation were always normal. It would be useful to generalize the findings to conditions where nonnormality is present.

Third, both the real and simulation analyses involve item parameters from a single test. The generalizability of the above results may be further supported by their replication with other tests. It might be suspected that the value of modeling distractors will vary somewhat depending on factors such as the difficulty of the linking items, and most certainly the number of distractor categories.

Fourth, this paper only considered one form of linking, namely the characteristic curve procedure originally presented by Baker (1993) and modified by Kim and Hanson (2002). Alternative linking procedures, such as concurrent calibration, might also be considered. One limitation of the characteristic curve procedure considered in this paper not shared by concurrent calibration is its lack of symmetry, implying that equating direction can influence the transformation performed. Generalization of the findings from this paper to other potential linking methods using the NRM would be useful.

Finally, although the current analysis is able to identify the relationship between linking precision under conditions of AC versus CO linking, the practical implications of linking imprecision on actual IRT applications is not clear. For example, if used for equating purposes, it may be that the amounts of linking imprecision observed are negligible in how they affect the score-to-score equating transformation. Moreover, other issues, such as the content representativeness of the linking items, are also important to consider in such applications, although were not made a part of the current analysis.

References

Baker, F.B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F.B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 16,* 239–251.

Baker, F.B., & Kim, S–H. (2004). *Item response theory: Parameter estimation techniques.* 2nd Ed. New York: Marcel Dekker.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113–141.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*, 381–409.

Center for Placement Testing (1998). *Mathematics Placement Test Form 98-X.* University of Wisconsin-Madison.

Dennis, J.E., & Schnabel, R.B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations.* Society for Industrial and Applied Mathematics. Philadelphia.

Drasgow, F., Levine, M.V., Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143–165.

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.

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

Kim, S.–H., & Cohen, A.S. (2002). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131–143.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling and linking.* 2nd Ed. New York: Springer.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog* (version 7) [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501–519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161–176.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Table 1. Nominal Response Model Parameters, Mathematics Placement Test

Item	Category Slopes					Category Intercepts				
	a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5
1	0.07	-0.22	-0.25	0.40	-0.01	0.38	-0.74	-0.90	1.35	-0.11
2	-0.62	-0.22	-0.17	1.27	-0.25	-0.94	-0.43	1.48	0.75	-0.84
3	-0.14	-0.05	-0.16	-0.37	0.73	-0.27	-0.27	-0.03	-1.08	1.66
4	-0.21	-0.60	0.39	0.55	-0.11	-0.10	-1.30	0.19	1.20	0.00
5	0.00	-0.12	0.36	-0.18	-0.05	0.06	-0.36	0.75	-0.48	0.03
6	-0.01	0.77	-0.61	0.19	-0.36	-0.79	1.21	0.33	0.39	-1.15
7	-0.34	-0.69	-1.07	1.32	0.78	0.35	-1.01	-2.05	1.60	1.13
8	-0.19	0.63	-0.22	-0.26	0.03	0.13	1.29	-0.69	-0.56	-0.15
9	-0.05	-0.55	-0.28	-0.29	1.18	-0.19	-1.30	-0.59	0.54	1.55
10	-0.12	-0.29	-0.23	0.50	0.13	-0.44	-0.43	0.15	0.21	0.52
11	0.62	-0.17	-0.43	-0.01	0.00	-0.04	0.91	-0.63	0.11	-0.36
12	-0.26	0.59	-0.20	-0.15	0.01	-0.30	0.72	0.33	-0.14	-0.59
13	-0.17	0.77	0.11	-0.47	-0.22	-0.63	1.83	0.81	-1.30	-0.72
14	-0.55	-0.45	0.98	-0.09	0.09	-1.18	-0.42	0.96	-0.51	1.15
15	-0.02	-0.46	0.85	-0.43	0.08	0.74	-0.92	0.97	-0.54	-0.23
16	-0.18	0.53	-0.45	0.13	-0.03	-0.67	0.83	-0.24	0.12	-0.06
17	0.13	-0.42	0.42	0.01	-0.15	0.30	-0.62	1.35	0.14	-1.19
18	0.88	-0.46	-0.45	-0.16	0.18	-0.08	-0.76	-0.19	0.21	0.80
19	-0.36	0.10	0.62	-0.21	-0.13	-0.99	0.47	1.03	-0.38	-0.14
20	0.10	-0.92	-0.24	-0.13	1.19	0.25	-2.24	0.36	0.64	1.00
21	-0.06	-0.20	-0.01	0.52	-0.26	-0.03	-0.51	-0.01	0.45	0.10
22	-0.04	-0.19	-0.19	0.48	-0.06	-0.15	0.30	0.10	0.46	-0.72
23	0.83	-0.27	-0.07	-0.70	0.22	2.07	-0.43	-0.12	-2.13	0.60
24	-0.53	-0.17	0.11	-0.03	0.64	-0.12	-0.53	0.27	-0.35	0.73
25	-0.22	0.48	-0.06	-0.15	-0.07	0.26	0.64	-0.65	-0.42	0.15
26	-0.09	0.65	-0.29	-0.37	0.08	0.14	0.76	-0.30	-0.50	-0.10
27	-0.31	-0.33	-0.11	-0.06	0.82	-1.11	-0.25	0.68	0.16	0.53
28	-0.43	-0.13	-0.21	0.87	-0.10	-0.78	0.51	-0.37	0.86	-0.24
29	0.91	-0.20	-0.38	-0.56	0.24	2.00	-0.26	-0.93	-1.30	0.49
30	-0.25	-0.54	0.17	0.62	0.02	0.84	-1.44	0.30	1.03	-0.74
31	0.87	-0.18	-0.45	-0.40	0.14	0.21	-0.01	0.42	-0.52	-0.08
32	0.67	-0.23	-0.10	-0.18	-0.17	0.18	-0.17	0.66	-0.61	-0.05
33	0.03	-0.13	0.50	-0.31	-0.08	-0.50	-0.21	1.45	0.02	-0.78
34	-0.41	-0.49	-0.24	0.95	0.17	-0.21	0.06	0.06	1.07	-1.00
35	-0.32	-0.44	0.65	0.22	-0.11	-0.24	-0.75	1.09	0.63	-0.75
36	0.75	-0.11	-0.36	-0.08	-0.18	0.24	-0.20	-0.34	0.64	-0.35

Table 2. Simulation Study, Root Mean Square Error in Recovery of Linking Intercept and Slope Parameters, Sample Size=1,000

True Linking Parameters		Linking Intercept (*I*)										Linking Slope (*S*)									
		Number of Linking Items										Number of Linking Items									
		2		5		10		15		20		2		5		10		15		20	
I	*S*	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO
-1.00	.5	.027	.088	.019	.033	.015	.022	.015	.020	.014	.019	.009	.025	.005	.013	.003	.007	.003	.005	.002	.004
-0.25	.5	.039	.112	.033	.049	.030	.035	.029	.031	.027	.031	.011	.026	.008	.015	.008	.011	.008	.010	.008	.009
0.00	.5	.041	.115	.032	.052	.032	.037	.030	.033	.031	.035	.016	.041	.013	.021	.012	.017	.012	.014	.012	.015
0.25	.5	.037	.110	.029	.042	.027	.033	.027	.028	.027	.026	.022	.055	.019	.025	.019	.021	.019	.022	.019	.021
1.00	.5	.026	.150	.010	.037	.004	.015	.002	.008	.002	.006	.064	.150	.070	.095	.068	.076	.068	.080	.068	.075
-1.00	1	.005	.026	.002	.008	.001	.004	.001	.002	.001	.002	.007	.016	.003	.007	.001	.004	.001	.003	.001	.002
-0.25	1	.006	.027	.004	.009	.003	.005	.002	.004	.002	.004	.005	.017	.003	.007	.002	.004	.002	.003	.001	.002
0.00	1	.006	.028	.003	.008	.002	.004	.002	.004	.002	.003	.006	.020	.004	.009	.002	.005	.002	.004	.002	.003
0.25	1	.006	.025	.003	.008	.002	.005	.002	.003	.002	.003	.007	.024	.005	.010	.004	.006	.003	.005	.003	.005
1.00	1	.007	.033	.002	.009	.001	.004	.001	.002	.000	.002	.020	.053	.018	.025	.015	.018	.015	.021	.015	.019
-1.00	2	.007	.033	.004	.011	.003	.006	.003	.004	.003	.004	.012	.034	.005	.014	.003	.007	.002	.005	.002	.004
-0.25	2	.007	.020	.004	.009	.003	.005	.003	.004	.003	.003	.008	.024	.003	.009	.002	.005	.001	.003	.001	.002
0.00	2	.006	.020	.004	.008	.003	.004	.003	.004	.003	.004	.007	.024	.003	.009	.001	.004	.001	.003	.001	.002
0.25	2	.007	.020	.004	.008	.003	.005	.003	.004	.003	.004	.008	.027	.003	.011	.002	.005	.001	.004	.001	.002
1.00	2	.007	.022	.004	.008	.003	.005	.003	.004	.002	.003	.009	.034	.004	.012	.002	.006	.001	.004	.001	.003
Average		.015	.055	.010	.020	.009	.013	.008	.011	.008	.010	.014	.038	.011	.019	.010	.013	.009	.012	.009	.011

AC: All Categories NRM used for linking. CO: Correct Category Only used for linking.

Table 3. Simulation Study, Linking Intercept and Slope Parameter Recovery Results: CO Linking versus 2PL Linking

	Linking Intercept (*I*)								Linking Slope (*S*)							
	Sample Size								Sample Size							
	250		500		1,000		3,000		250		500		1,000		3,000	
#Items/Method	CO	2PL	CO	2PL	CO	2PL	CO	2PL	CO	2PL	CO	2PL	CO	2PL	CO	2PL
2	.394	.377	.128	.120	.055	.061	.016	.015	.356	.373	.100	.094	.038	.038	.010	.009
5	.196	.200	.043	.044	.020	.021	.006	.007	.132	.141	.050	.045	.019	.020	.005	.005
10	.135	.128	.025	.028	.013	.014	.003	.003	.082	.088	.036	.034	.013	.013	.003	.003
15	.122	.116	.022	.023	.011	.012	.003	.003	.062	.059	.033	.032	.012	.012	.003	.003
20	.124	.125	.020	.020	.010	.011	.002	.003	.056	.056	.031	.032	.011	.012	.003	.003

CO: Correct Category Only used for linking.

Table 4. Real Data Study, Root Mean Square Differences of Non-Linking Common Item Parameter Estimates, Sample Size=1,000

Equating group standardized test score distribution		Category Intercepts										Category Slopes									
		Number of Linking Items										Number of Linking Items									
		2		5		10		15		20		2		5		10		15		20	
Mean	SD	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO	AC	CO
-1	.5	.146	.302	.130	.202	.127	.147	.130	.133	.129	.132	.156	.269	.122	.272	.117	.158	.116	.128	.118	.128
0	.5	.127	.139	.103	.121	.102	.112	.098	.104	.100	.102	.027	.060	.019	.049	.019	.031	.016	.021	.016	.019
1	.5	.110	.130	.101	.120	.099	.107	.100	.103	.099	.105	.083	.110	.072	.103	.067	.080	.066	.073	.065	.072
-1	1	.116	.169	.098	.125	.087	.108	.085	.095	.083	.094	.123	.256	.084	.152	.069	.119	.063	.089	.060	.092
0	1	.056	.067	.049	.058	.048	.049	.048	.051	.047	.050	.019	.029	.016	.018	.014	.014	.014	.017	.014	.015
1	1	.045	.053	.040	.048	.040	.044	.039	.041	.039	.040	.027	.041	.022	.031	.020	.025	.019	.021	.019	.021
-1	2	.043	.050	.038	.044	.037	.039	.036	.038	.037	.038	.042	.048	.037	.043	.033	.036	.032	.035	.033	.034
0	2	.020	.026	.018	.022	.018	.019	.017	.017	.017	.018	.012	.023	.010	.017	.009	.010	.009	.010	.009	.009
1	2	.026	.030	.024	.027	.023	.024	.023	.024	.023	.024	.015	.025	.014	.018	.014	.014	.014	.014	.013	.014
Average		.077	.107	.067	.085	.064	.072	.064	.067	.064	.067	.056	.096	.044	.078	.040	.054	.039	.045	.039	.045

AC: All Categories NRM used for linking. CO: Correct Category Only used for linking.

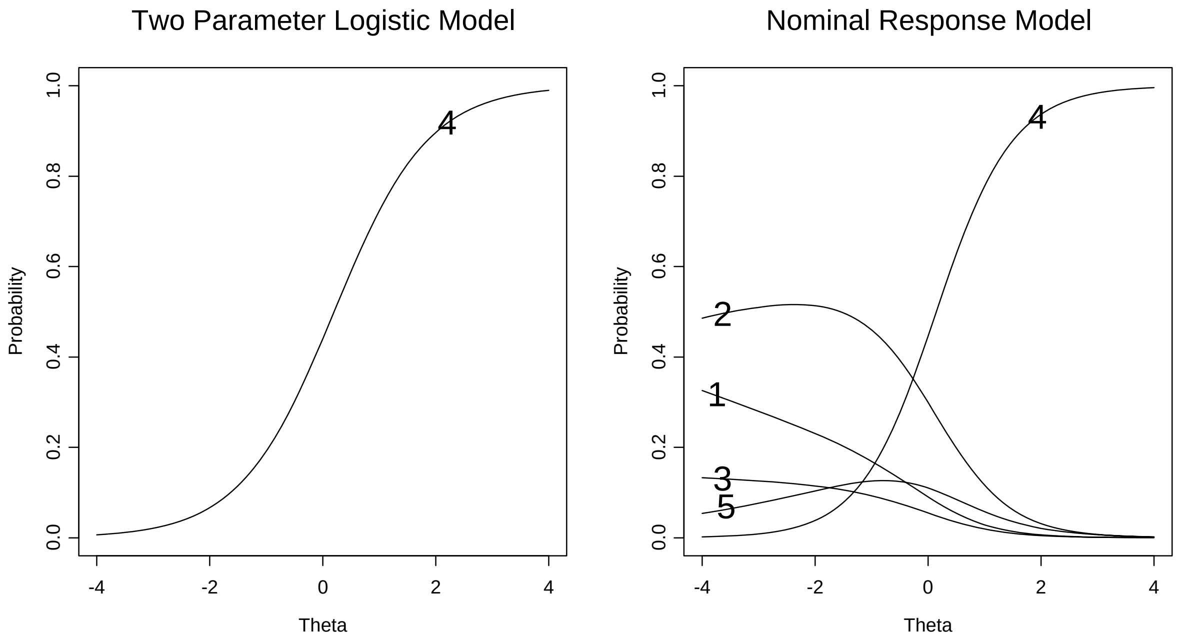

Figure 1. Example 2PL and Nominal Response Model Items

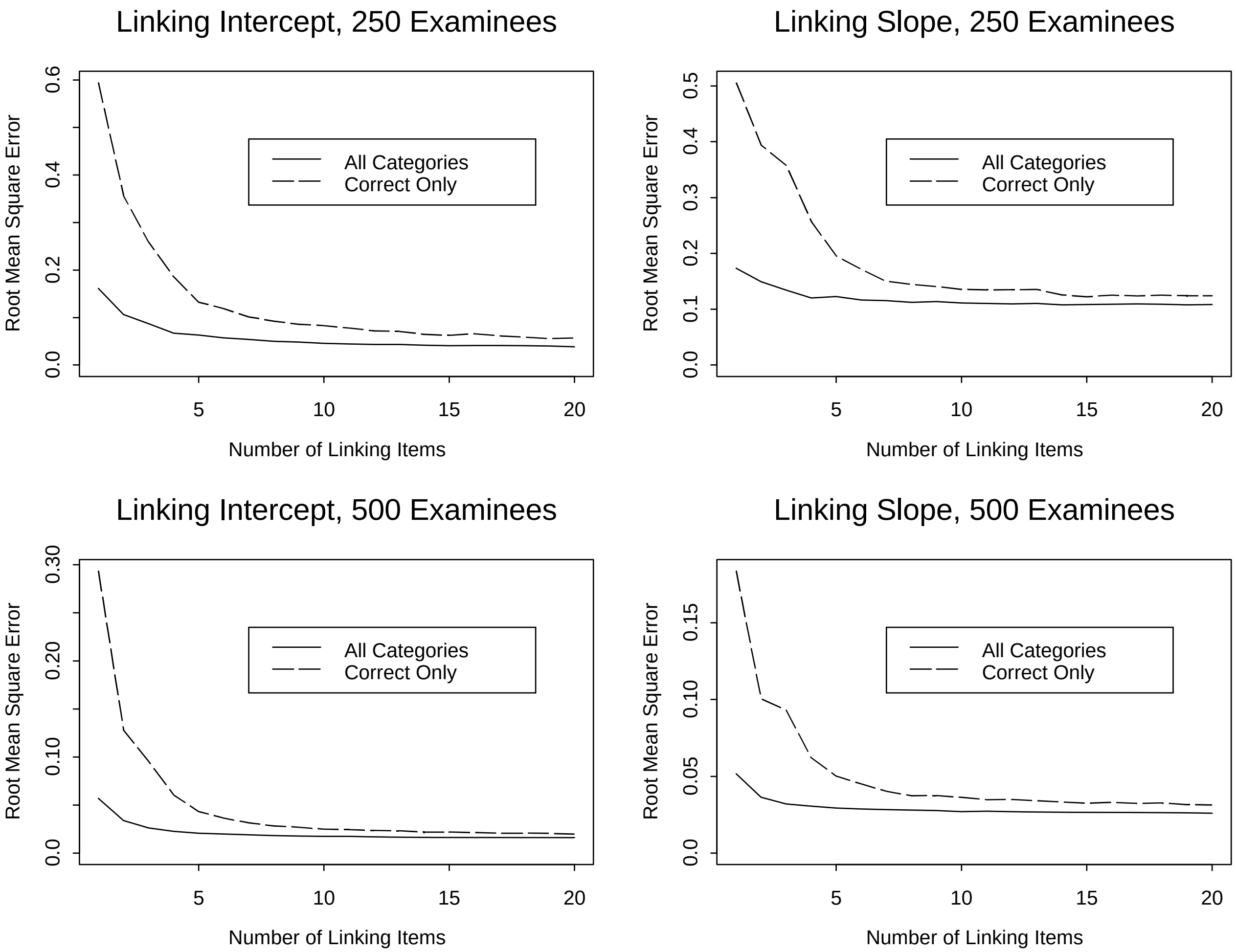

Figure 2a. Recovery of Linking Intercept and Slope Parameters under Correct-Only and All-Category Linkings, Sample Sizes=250, 500

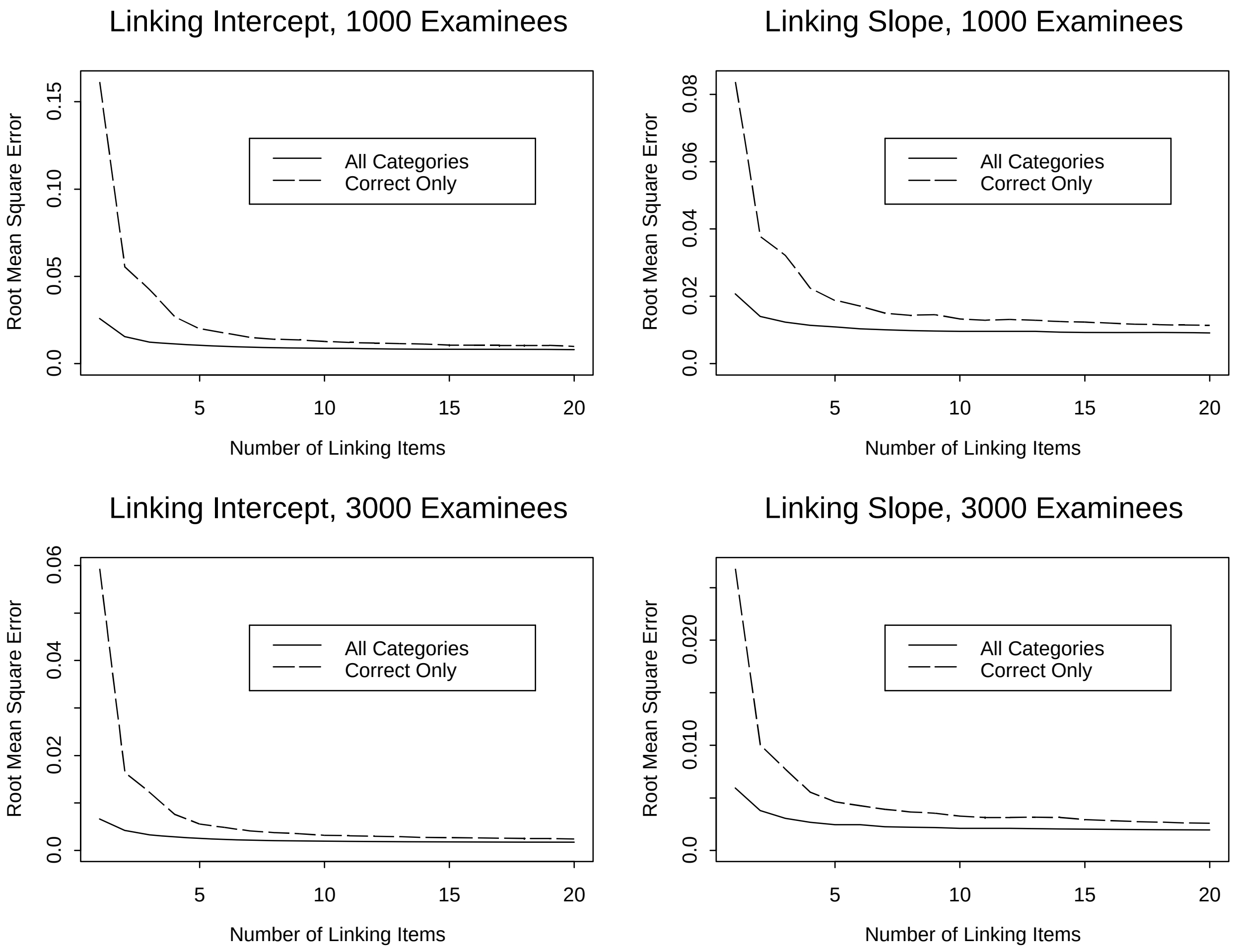

Figure 2b. Recovery of Linking Intercept and Slope Parameters under Correct-Only and All-Category Linkings, Sample Sizes=1000, 3000

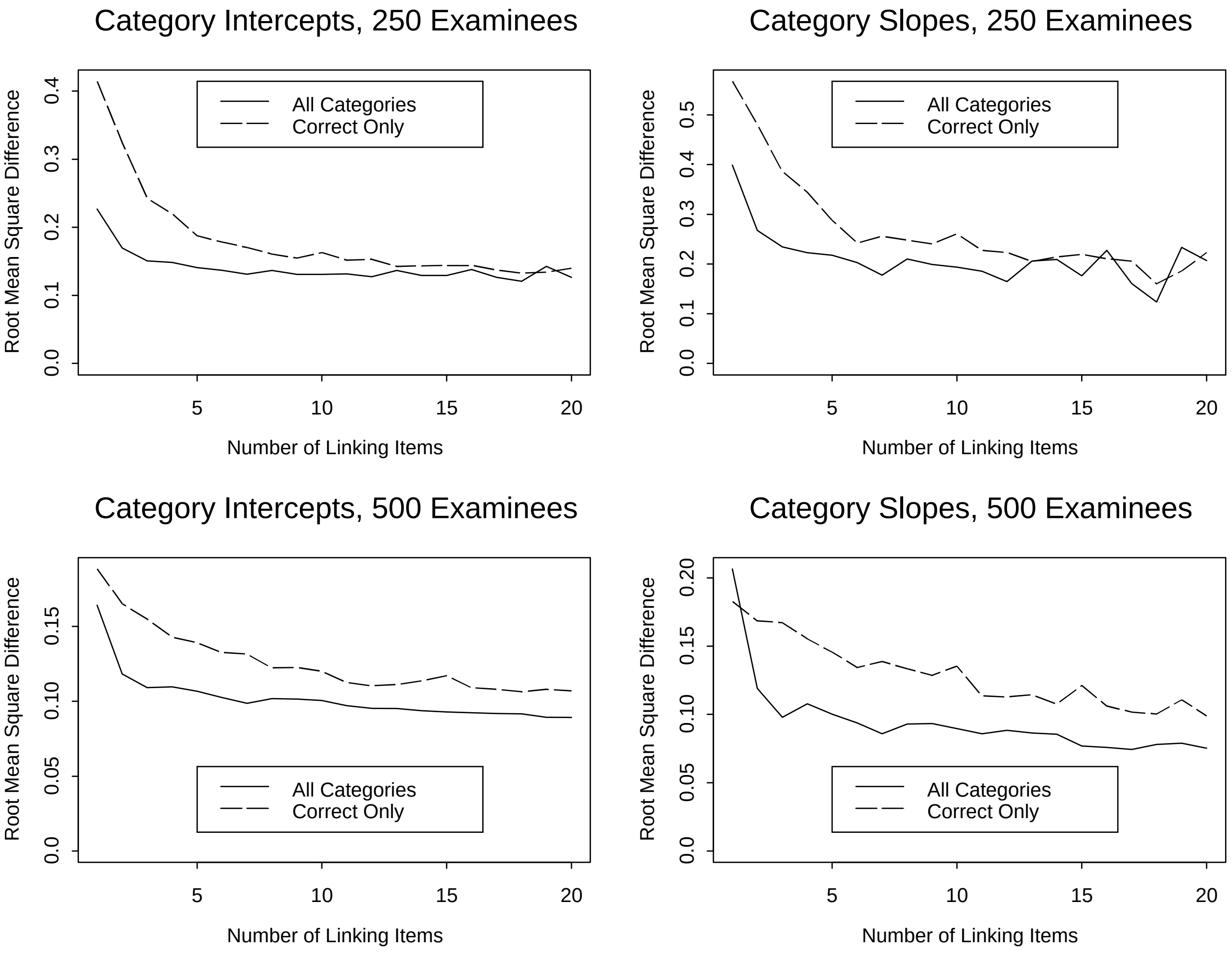

Figure 3a. Equivalence of Item Category Intercepts and Slopes, Non-Linking Common Items, Sample Sizes=250, 500

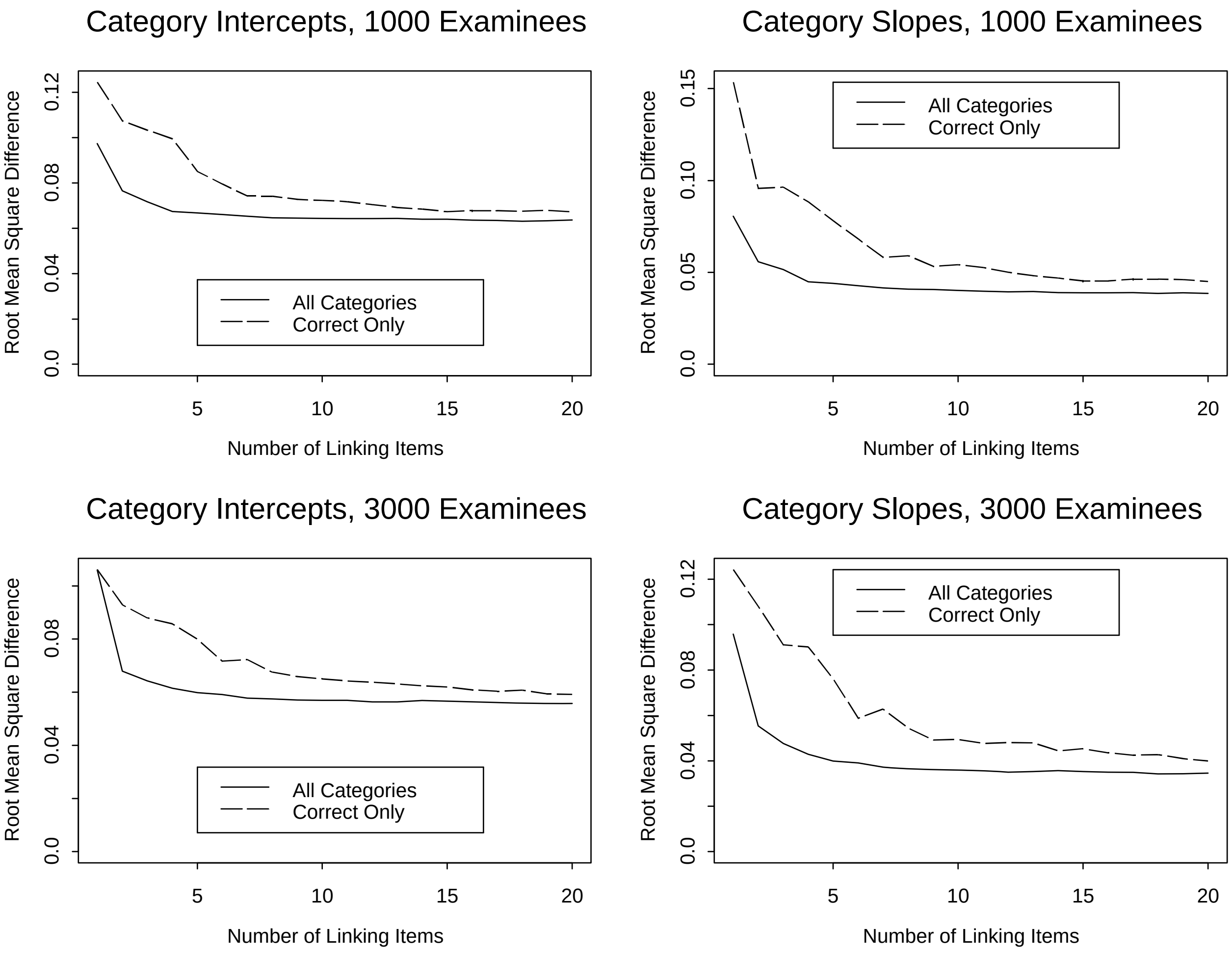

Figure 3b. Equivalence of Item Category Intercepts and Slopes, Non-Linking Common Items, Sample Sizes=1000, 3000