A SIMULATION STUDY ON THE PERFORMANCE OF FOUR MULTIDIMENSIONAL IRT SCALE LINKING METHODS

By

YOUHUA WEI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Dr. James J. Algina, my supervisory committee chair, for providing valuable guidance and support. I would also like to thank other committee members, Dr. M. David Miller, Dr. Walter L. Leite, and Dr. Zhihui Fang, for their time and effort on this project.

I thank my parents and my brothers and sisters for their continuous and unconditional support and encouragement. Finally, I thank my wife, Yan Zhang, for her love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

A SIMULATION STUDY ON THE PERFORMANCE OF
FOUR MULTIDIMENSIONAL IRT SCALE LINKING METHODS

By

Youhua Wei

August 2008

Chair: James J. Algina
Major: Research and Evaluation Methodology

Scale linking is the process of developing the connection between scales of two or more sets of parameter estimates obtained from separate test calibrations. It is the prerequisite for many applications of IRT, such as test equating and differential item functioning analysis. Unidimensional scale linking methods have been studied and applied frequently over the past two decades. The development of multidimensional linking methods is at the infancy stage and more research is needed to obtain definitive results.

As an extension of previous research, the purpose of this study was to use simulated data to evaluate the performance of four multidimensional IRT scale linking methods, the direct method, equated function method, test characteristic function method, and item characteristic function method, under various testing conditions, which include different test structures, test lengths, sample sizes, and ability distributions. There were one hundred and ninety-two experimental conditions in this study and five hundred replications were conducted for each of the conditions. The linking performance evaluation was based on the differences between the item parameter estimates for base group and the transformed item parameter estimates for the equated group

across the test items. The mean and standard deviation of the differences across the 500 replications were computed to examine the accuracy and stability of the four linking methods.

Our results indicate that for approximate simple test structure, each of the four linking methods worked approximately equally well under all testing conditions. The results also suggest that for complex test structure: (a) The equated function method did not work well under any testing conditions, (b) the performance of other three linking methods depended on other testing conditions including sample size, test length, and ability distribution difference between groups, and (c) the direct method was the best linking procedure for most testing conditions. In addition, the study shows that the item parameter values influenced the linking performance. Under most of the testing conditions, the linking results for the discrimination parameter tended to be less accurate and less stable when the item parameter had extreme values. The linking accuracy for the difficulty parameter was not dependent on the item parameter values. The linking stability for the difficulty parameter depended on the item parameter values only when the sample size was large. Then, the linking results were less stable when the item parameter had extreme values.

CHAPTER 1
INTRODUCTION

Suppose a set of test items is administered to non-equivalent groups of examinees and item response theory (IRT) is used to estimate the item parameters for each of the groups. The parameter estimates will be on different scales because the metric defined by each separate calibration is different (Stocking & Lord, 1983). Specifically, IRT parameter estimation procedures often scale the ability for each group with mean of 0 and standard deviation of 1, although the actual ability distributions of the two groups may be different (Kolen & Brennan, 2004). Therefore, to compare the parameter estimates from different IRT calibrations, they should be transformed on the same scale. Scale linking is the process of developing the connection between scales of two or more sets of parameter estimates obtained from separate test calibrations. The objective is to establish a common metric for all sets of parameter estimates.

Scale linking is an important issue in psychometrics, and many applications of IRT require that item parameter estimates from independent calibrations be expressed on the common metric, including test equating and differential item functioning (DIF) (Stocking & Lord, 1983). Based on Kolen and Brennan (2004), equating is "a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably." (p. 2), and linking refers to relating scores on tests which are not built to the same content or statistical specifications. Different terminologies have been used to describe the process of establishing relationship between scores on two or more tests (for a complete review, see Kolen, 2004a, 2004b). Scale linking is used in this study to refer to the process of linking different scales rather than the process of linking test scores. However, scale linking is the prerequisite for establishing the connection between different test scores. Therefore, scale linking is an important step in test

equating (Cook & Eignor, 1991; Kolen & Brennan, 2004) and satisfactory equating results require successful scale linking.

If different groups of examinees have different probabilities of success on an item after they have been matched on the ability of interest, the item has differential functioning. In IRT, DIF is defined as the differences in the model parameters for the comparison groups (Clauser & Mazor, 1998). The item parameters for different groups should be compared only after they are placed on a common metric. Therefore, DIF identification depends heavily on the quality of scale linking. Some procedures have been developed to detect DIF by improving scale linking (Candell & Drasgow, 1988; Lautenschlager & Park, 1988; Lautenschlager, Flaherty, & Park, 1994; Park & Lautenschlager, 1990).

In addition to psychometrics, scale linking is also very important to educational and psychological studies. Multi-group confirmatory factor analysis or mean and covariance structure analysis has been increasingly used to compare constructs across different groups (for a comprehensive review, see Vandenberg, 2000) and some unresolved issues are closely related to the difficulty of linking scales across groups (Millsap, 2005). Therefore, successful scale linking has the potential to produce satisfactory comparison studies on psychological constructs across different groups.

In sum, scale linking is very important for educational measurement to be fair and objective for different groups of examinees. Unidimensional scale linking methods have been studied and applied frequently over the past two decades (for more information, see Kolen & Brennan, 2004; Yen & Fitzpatrick, 2006). The development of multidimensional linking methods (Davey, Oshima, & Lee, 1996; Hirsch, 1988, 1989; Li, 1997; Li & Lissitz, 2000; Min, 2003; Oshima, Davey, & Lee, 2000) is just at the infancy stage and more research is needed to

obtain definitive results (Yen & Fitzpatrick, 2006). In this chapter, unidimensional and multidimensional models and linking methods are reviewed and the purpose of the current study is presented.

Unidimensional IRT Models

Logistic Model

The three-parameter logistic (3PL) model (see Hambleton & Swaminathan, 1985; Lord, 1980) assumes that the probability of a correct answer to a dichotomously scored item j by an examinee with ability θ_i is

$$\begin{aligned} & P_{ij}\left(x_{ij}=1\middle|\theta_i;a_j,b_j,c_j\right) \\ & = c_j+\left(1-c_j\right)\frac{e^{a_j\left(\theta_i-b_j\right)}}{1+e^{a_j\left(\theta_i-b_j\right)}} \\ & = c_j+\left(1-c_j\right)\frac{1}{1+e^{-[a_j\left(\theta_i-b_j\right)]}}, \end{aligned} \tag{1-1}$$

where x_{ij} is the item response (0 or 1) for person i on test item j,

a_j is the item discrimination parameter,

b_j is the item difficulty parameter, and

c_j is the guessing parameter or the pseudo-chance score level, representing the probability of correct response when the ability assessed by the item is very low.

Sometimes the 3PL model is expressed as

$$P_{ij}\left(x_{ij}=1\middle|\theta_i;a_j,b_j,c_j\right)=c_j+\left(1-c_j\right)\frac{1}{1+e^{-[Da_j\left(\theta_i-b_j\right)]}}, \tag{1-2}$$

with D=1.701, so that a normal ogive model item characteristic curve (ICC) and a logistic model ICC with the same item parameters are almost identical.

If c_j is 0, the 3PL model becomes two-parameter logistic (2PL) model:

$$P_{ij}\left(x_{ij}=1\middle|\theta_i;a_j,b_j\right)=\frac{1}{1+e^{-[a_j(\theta_i-b_j)]}}. \tag{1-3}$$

For 2PL model, if a_j is 1, it becomes one-parameter logistic (1PL) model or Rasch model:

$$P_{ij}\left(x_{ij}=1\middle|\theta_i;b_j\right)=\frac{1}{1+e^{-(\theta_i-b_j)}}. \tag{1-4}$$

Normal Ogive Model

There are also three normal ogive models or cumulative normal distribution models in IRT: one parameter model:

$$P_{ij}\left(\theta_i;b_j\right)=\int_{-\infty}^{(\theta_i-b_j)}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(t^2)}dt; \tag{1-5}$$

two parameter model:

$$P_{ij}\left(x_{ij}=1\middle|\theta_i;a_j,b_j\right)=\int_{-\infty}^{a_j(\theta_i-b_j)}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(t^2)}dt; \tag{1-6}$$

and three parameter model:

$$P_{ij}\left(x_{ij}=1\middle|\theta_i;a_j,b_j,c_j\right)=c_j+(1-c_j)\int_{-\infty}^{a_j(\theta_i-b_j)}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(t^2)}dt. \tag{1-7}$$

Many IRT models have been developed for test items that are polytomously scored using ordered categories, including graded response model (Samejima,1969), partial credit model (Masters, 1982), generalized partial credit model (Muraki, 1992), rating scale model (Andrich, 1978), and nominal response model (Bock, 1972).

Unidimensional IRT Scale Linking

Scale Transformation

The IRT parameter estimates produced from independent calibrations using data obtained from different groups of examinees are often on different metrics. Lord (1980) demonstrated that

the relationship between the metrics of any two independent item calibrations is linear. Therefore, a linear equation can be used to transform the IRT parameters on scale E (representing the linked scale or equated scale) to scale F (representing the base scale). For person i and item j,

$$\theta^*_{F_i} = A\theta_{E_i} + B, \tag{1-8}$$

$$a^*_{F_j} = \frac{a_{E_j}}{A}, \tag{1-9}$$

$$b^*_{F_j} = Ab_{E_j} + B, \tag{1-10}$$

$$c^*_{F_j} = c_{E_j}, \tag{1-11}$$

where $\theta^*_{F_i}$, $a^*_{F_j}$, $b^*_{F_j}$, and $c^*_{F_j}$ represent the transformed values from the linked scale to the base scale. A is the slope and B is the intercept. The constants A and B can be expressed as

$$A = \frac{a_{E_j}}{a^*_{F_j}}, \tag{1-12}$$

$$B = \theta^*_{F_i} - A\theta_{E_i} = b^*_{F_j} - Ab_{E_j}. \tag{1-13}$$

A and B can also be expressed for any two individuals i and i^* or two items j and j^*:

$$A = \frac{\theta_{F_i} - \theta_{F_{i^*}}}{\theta_{E_i} - \theta_{E_{i^*}}} = \frac{b_{F_j} - b_{F_{j^*}}}{b_{E_j} - b_{E_{j^*}}}, \tag{1-14}$$

$$B = b_{F_j} - Ab_{E_j} = \theta_{F_i} - A\theta_{E_i}, \tag{1-15}$$

or expressed for groups of items or examinees (see Kolen & Brennan, 2004):

$$A = \frac{\sigma(b_F)}{\sigma(b_E)} = \frac{\sigma(\theta_F)}{\sigma(\theta_E)} = \frac{\mu(a_E)}{\mu(a_F)}, \tag{1-16}$$

$$B = \mu(b_F) - A\mu(b_E) = \mu(\theta_F) - A\mu(\theta_E). \tag{1-17}$$

The $P_{ij}\left(\theta_{E_i}\right)$ value for the original parameters on scale E will be the same as the $P_{ij}\left(\theta^*{}_{F_i}\right)$ value for the transformed parameters on scale F as demonstrated by

$$
\begin{aligned}
& P_{ij}\left(\theta^*{}_{F_i}\right) \\
& = c^*{}_{F_j} + \left(1 - c^*{}_{F_j}\right)\frac{1}{1+e^{-\left[Da^*{}_{F_j}\left(\theta^*{}_{F_i} - b^*{}_{F_j}\right)\right]}} \\
& = c_{E_j} + \left(1 - c_{E_j}\right)\frac{1}{1+e^{-\left[D\frac{a_{E_j}}{A}\left(A\theta_{E_i} + B - Ab_{E_j} - B\right)\right]}} \\
& = c_{E_j} + \left(1 - c_{E_j}\right)\frac{1}{1+e^{-\left[Da_{E_j}\left(\theta_{E_i} - b_{E_j}\right)\right]}} \\
& = P_{ij}\left(\theta_{E_i}\right).
\end{aligned}
$$

Therefore, the logistic function is invariant under a linear transformation of item and ability parameters. Most of the unidimensional IRT scale linking methods are based on this important feature.

Scale Linking

In practice, both test item parameters and examinees' ability parameters need to be estimated and the ability estimates are often scaled to have means of 0 and standard deviations of 1. Parameter estimates obtained from different groups of examinees are often on different scales due to nonequivalence of the groups even though all ability estimates are scaled with means of 0 and standard deviations of 1. Therefore, some data collection procedures are required to establish the connection between different scales by using the linear transformations mentioned above. In test equating, three data collection designs are often used, including random groups design, single group design, and common-item nonequivalent groups design. The IRT parameter estimates for the first and second designs are assumed to be on the same scale because of the randomly equivalent groups of examinees and single group of examinees (Kolen & Brennan, 2004) if random sampling errors are ignored. For the third design, the parameter estimates are

assumed to be on different scales due to the nonequivalent groups. The third design is the most often used equating design (Kolen & Brennan, 2004) and it is very similar to the design used for exploring DIF. Two approaches have been used to establish a common scale for parameter estimates for this design. One is to estimate parameters for all items on both test forms together. This method is often called concurrent calibration (Wingersky & Lord, 1984). Both BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and MULTILOG (Thissen, 1991) have the function of simultaneously obtaining parameter estimates for two test forms and two groups on the same scale. The second approach is to link the two scales by using the parameter estimates for the common items. This study will focus on the second approach. The following IRT linking methods have been developed to establish a common metric for parameter estimates.

Mean/sigma method. This method (Marco, 1977) uses the means and standard deviations of the *b* parameter estimates for the common items to calculate the constants A and B in the linear transformation equation:

$$A = \frac{\hat{\sigma}(b_F)}{\hat{\sigma}(b_E)},\ B = \hat{\mu}(b_F) - A\hat{\mu}(b_E). \tag{1-18}$$

Mean/mean method. This method (Loyd & Hoover, 1980) uses the means of *a* parameter estimates for the common items to calculate A and the means of *b* parameter estimates for the common items to calculate B in the transformation equation:

$$A = \frac{\hat{\mu}(a_E)}{\hat{\mu}(a_F)},\ B = \hat{\mu}(b_F) - A\hat{\mu}(b_E). \tag{1-19}$$

Item response function method. In this procedure (Haebara, 1980), the constants A and B are estimated by minimizing the sum of the squared difference between the item characteristic curves for the common items over examinees:

$$H_{diff} = \sum_i \sum_j \left[P_{ij}\left(\theta_{F_i}; \hat{a}_{F_j}, \hat{b}_{F_j}, \hat{c}_{F_j}\right) - P_{ij}\left(\theta_{F_i}; \frac{\hat{a}_{E_j}}{A}, A\hat{b}_{E_j} + B, \hat{c}_{E_j}\right)\right]^2 . \qquad (1\text{-}20)$$

Test response function method. The constants A and B are estimated by minimizing the sum of the squared difference between the test characteristic curves for the common items for examinees (Stocking & Lord, 1983):

$$SL_{diff} = \sum_i \left[\sum_j P_{ij}\left(\theta_{F_i}; \hat{a}_{F_j}, \hat{b}_{F_j}, \hat{c}_{F_j}\right) - \sum_j P_{ij}\left(\theta_{F_i}; \frac{\hat{a}_{E_j}}{A}, A\hat{b}_{E_j} + B, \hat{c}_{E_j}\right)\right]^2 . \qquad (1\text{-}21)$$

Item response function method and test response function method are often referred as the characteristic curve methods (Stocking & Lord, 1983). Specifically, the former is called item characteristic method and the latter test characteristic curve method.

Minimum χ^2 method. This method (Divgi, 1985) combines information of each item's parameter estimates and the variance-covariance matrix of sampling errors for each item from the item parameter estimation procedure. The constants A and B are estimated by minimizing the following quadratic function:

$$\chi^2 = \sum_j \left[\hat{a}_{F_j} - \frac{\hat{a}_{E_j}}{A}, \hat{b}_{F_j} - \left(A\hat{b}_{E_j} + B\right)\right] \left[\hat{\Sigma}_{F_j} + \hat{\Sigma}^*_{F_j}\right]^{-1} \left[\hat{a}_{F_j} - \frac{\hat{a}_{E_j}}{A}, \hat{b}_{F_j} - \left(A\hat{b}_{E_j} + B\right)\right]' , \qquad (1\text{-}22)$$

where $\hat{\Sigma}_{F_j}$ is the estimated variance-covariance matrix of the sampling errors for the item parameter estimates for item *j* on the F scale and $\hat{\Sigma}^*_{F_j}$ is the estimated variance-covariance matrix of the sampling errors for the item parameter estimates for item *j* which are transformed from the E scale to the F scale.

Comparison studies have been conducted for these methods with dichotomous IRT model. Based on a comprehensive literature review (Kolen & Brennan, 2004): (a) The

characteristic curve methods produced more stable and accurate results than the mean/mean and mean/sigma methods, (b) the mean/mean method was more stable than the mean/sigma method, (c) the concurrent calibration method yielded more accurate results than the test characteristic curve method for a small number of common items and both procedures had the similar accuracy for a larger number of common items, and (d) the concurrent calibration method might be less robust to violations of the IRT assumptions than characteristic curve methods.

These methods have been extended to link scales with polytomous IRT models. For example, Cohen and Kim (1998) extended mean/mean and mean/sigma methods to the graded response model and Kolen and Brennan (2004) suggested using mean/mean and mean/sigma methods for the generalized partial credit model. Baker (1992) generalized the test response function method to the graded response model and Baker (1993) used the item response function for the nominal response model. Kim and Cohen (1995) tried the minimum χ^2 method for the graded response model.

There are also some comparison studies for these methods with polytomous IRT models. A simulation study (Cohen & Kim, 1998) comparing the mean/mean method, mean/sigma method, weighted mean/sigma method, test response function method, and minimum χ^2 method for the graded response model found that all methods produced similar results. Another simulation study (Kim & Cohen, 2002) comparing the test response function method and the concurrent calibration method for the graded response model found that the concurrent calibration was relatively more accurate.

Multidimensional IRT Models

Logistic Model

Unidimensional IRT models appear to be adequate for scaling achievement test items in most practical situations (Yen & Fitzpatrick, 2006). However, it is reasonable to believe that the performance of examinees on some test items depends on more than one trait or ability and some consequences of applying unidimensional models to multidimensional data have been identified (see Yen & Fitzpatrick, 2006). The number of dimensions necessary to model the test item responses depends not only on the number of ability dimensions and the level on those dimensions exhibited by the examinees but also on the number of skills to which the test items are sensitive (Reckase, 1997a). Therefore, multidimensionality can occur in different ways depending on the interaction between a specific group of examinees and certain set of test items.

There are two types of multidimensional IRT (MIRT) models for dichotomously scored item response data: the compensatory model and the noncompensatory model. In the compensatory model, a low θ value on one dimension can be compensated for by a high θ value on another dimension (McKinley & Reckase, 1983; Reckase, 1997a). In the noncompensatory model, an increase in the θ value on one dimension cannot compensate for a lower value on another dimension (Simpson, 1978). Since estimation programs and linking methods have not been well developed for noncompensatory model, the most often used compensatory model is discussed and used in this study.

The compensatory multidimensional three-parameter logistic (M3PL) model is a direct generalization of the unidimensional 3PL model (Reckase, 1997a):

$$
\begin{aligned}
&P(x_{ij}=1|\boldsymbol{\theta}_i;\mathbf{a}_j,d_j,c_j)\\
&=c_j+(1-c_j)\frac{e^{(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}}{1+e^{(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}}\\
&=c_j+(1-c_j)\frac{1}{1+e^{-(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}},
\end{aligned}
\tag{1-23}
$$

where $P(x_{ij}=1|\boldsymbol{\theta}_i;\mathbf{a}_j,d_j,c_j)$ is the probability of a correct response ($x_{ij}=1$) for person i on test item j,

x_{ij} is the item response (0 or 1) for person i on test item j,

$\mathbf{a}_j$ is the vector of item discrimination parameters,

d_j is the scalar parameter related to the difficulty of the item,

c_j is the lower asymptote or guessing parameter, and

$\boldsymbol{\theta}_i$ is the vector of ability parameters for person i.

This model can be expressed in the following scalar form:

$$
\begin{aligned}
&P(x_{ij}=1|\theta_i;a_j,d_j,c_j)\\
&=c_j+(1-c_j)\frac{e^{\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}}{1+e^{\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}}\\
&=c_j+(1-c_j)\frac{1}{1+e^{-\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}},
\end{aligned}
\tag{1-24}
$$

where m is the number of dimensions. When c_j is 0, the compensatory M3PL model becomes the compensatory multidimensional two-parameter logistic (M2PL) model (McKinley & Reckase, 1983):

$$
P(x_{ij}=1|\boldsymbol{\theta}_i;\mathbf{a}_j,d_j)=\frac{e^{(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}}{1+e^{(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}}=\frac{1}{1+e^{-(\mathbf{a}_j'\boldsymbol{\theta}_i+d_j)}}. \tag{1-25}
$$

This model can also be expressed as the following scalar form:

$$P(x_{ij}=1|\theta_i;a_j,d_j)=\frac{e^{\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}}{1+e^{\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}}=\frac{1}{1+e^{-\left(\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j\right)}}. \tag{1-26}$$

Compared with unidimensional IRT models, multidimensional discrimination and ability parameters are described in the form of vectors instead of scalars. If the $\boldsymbol{\theta}_i$ dimensions are orthogonal, the observed correlations among the item scores will be accounted for by the $\mathbf{a}_j$ parameters. Otherwise, the item correlations will reflect both the $\mathbf{a}_j$ parameters and correlated $\boldsymbol{\theta}$ dimensions. In MIRT, the probability of a correct response to an item depends on multidimensional ability and is defined as an item characteristic surface (ICS). Assuming orthogonal axes of dimensions in the surface, an item *j* can be described by the following three characteristics (Ackerman, 1994; Reckase, 1985; Reckase, 1997a; Reckase & McKinley, 1991): multidimensional discrimination ($MDISC_j$):

$$MDISC_j=\sqrt{\sum_{k=1}^{m}a_{jk}^2}, \tag{1-27}$$

which is the discrimination power of the item for the most discriminating combination of dimensions; multidimensional difficulty ($MDIFF_j$):

$$MDIFF_j=\frac{-d_j}{MDISC_j}, \tag{1-28}$$

which, similar to the difficulty parameter in unidimensional model, is the distance from the origin of the θ space to the point of steepest slope in a direction from the origin; and direction (α_{jk}) of the greatest slope from the origin:

$$\alpha_{jk}=\arccos\frac{a_{jk}}{MDISC_j}, \tag{1-29}$$

which is the angle that the line from the origin of the space to the point of steepest slope makes with the *k*th axis for the item.

Normal Ogive Model

By adapting Thurstone's multiple factor model (1947) to dichotomous item response data, Bock and Aitkin (1981) proposed a multidimensional normal ogive model by firstly assuming that an unobserved continuous response variable, y_{ij}, for person i and item j is a linear combination of *m* latent variables, $\boldsymbol{\theta}$, weighted by the factor loadings, $\boldsymbol{\alpha}$:

$$y_{ij} = \alpha_{j1}\theta_{1i} + \alpha_{j2}\theta_{2i} + \cdots + \alpha_{jm}\theta_{mi} + \delta_{ij}, \tag{1-30}$$

where $\boldsymbol{\theta} \sim N(\mathbf{0},\mathbf{I})$, $\mathbf{y} \sim N(\mathbf{0},\mathbf{1})$, and $\boldsymbol{\delta} \sim N(\mathbf{0},\boldsymbol{\sigma}_j^2)$ (Note that the αs in Equation 1-30 are not the αs in Equation 1-29). It is assumed that there is an underlying process which generates a correct observed response, $x_{ij} = 1$, when y_{ij} equals or exceeds a threshold, γ_j, and produces an incorrect observed response, $x_{ij} = 0$, otherwise. Then the probability of obtaining a correct item score is

$$\begin{aligned} & P_{ij}\left(x_{ij} = 1 \middle| \theta_i; \alpha_j, \gamma_j\right) \\ &= \frac{1}{\sqrt{(2\pi)}\sigma_j} \int_{\gamma_j}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{y_{ij} - \sum_k^m \alpha_{jk}\theta_{ki}}{\sigma_j}\right)^2\right] dy_{ij} \\ &= \Phi\left(-\frac{\gamma_j - \sum_k^m \alpha_{jk}\theta_{ki}}{\sigma_j}\right) \\ &= \Phi_j(\theta_i), \end{aligned} \tag{1-31}$$

where $\sigma_j^2 = 1 - \sum_{k=1}^{m} \alpha_{jk}^2$. This is a compensatory model because greater ability on one dimension can make up for lesser ability on other dimensions. This model can be reparameterized to

produce similar parameters in multidimensional logistic model (Bock, Gibbons, & Muraki 1988; Muraki & Engelhard, 1985) by

$$\begin{aligned} & P_{ij}\left(x_{ij}=1\middle|\boldsymbol{\theta}_i;\mathbf{a}_j,d_j\right) \\ & =\frac{1}{\sqrt{2\pi}}\int_{-Z(\theta_i)}^{\infty}\exp\left(-\frac{t^2}{2}\right)dt \\ & =\int_{-z(\theta_i)}^{\infty}\phi(t)dt \\ & =\Phi_j(\theta_i), \end{aligned} \tag{1-32}$$

where

$$Z(\theta_i)=\sum_{k=1}^{m}a_{jk}\theta_{ik}+d_j=\mathbf{a}_j^{'}\boldsymbol{\theta}+d_j, \tag{1-33}$$

$$a_{jk}=\frac{\alpha_{jk}}{\sigma_j}, \tag{1-34}$$

$$d_j=-\frac{\gamma_j}{\sigma_j}. \tag{1-35}$$

It can also be shown that

$$\alpha_{jk}=\frac{a_{jk}}{q_i}, \tag{1-36}$$

$$\gamma_j=-\frac{d_j}{q_j}, \tag{1-37}$$

with

$$q_j^2=1+\sum_{k=1}^{m}a_{jk}^2. \tag{1-38}$$

When θs are correlated with covariance matrix $\boldsymbol{\Phi}$**,** it can be shown that

$$a_{jk}=\frac{\alpha_{jk}}{\sqrt{1-\boldsymbol{\alpha}`\boldsymbol{\Phi}\boldsymbol{\alpha}}}, \tag{1-39}$$

$$d_j = -\frac{\gamma_j}{\sqrt{1-\boldsymbol{\alpha}`\boldsymbol{\Phi}\boldsymbol{\alpha}}}, \tag{1-40}$$

$$\alpha_{jk} = \frac{a_{jk}}{\sqrt{1+\mathbf{a}`\boldsymbol{\Phi}\mathbf{a}}}, \tag{1-41}$$

$$\gamma_j = -\frac{d_j}{\sqrt{1+\mathbf{a}`\boldsymbol{\Phi}\mathbf{a}}}, \tag{1-42}$$

where $\boldsymbol{\alpha}$ is vector of factor loading for item *j* and **a** is the vector of discrimination for item *j*.

Multidimensional IRT models for polytomously scored test items have also been developed, including multidimensional logistic models and multidimensional normal ogive models (Kelderman, 1997; Muraki, 1999; Muraki & Carlson, 1995).

Multidimensional IRT Scale Linking

The multidimensional scale linking is more complicated than the unidimensional scale linking because it involves the transformations of scale locations, variances, and covariances of several ability dimensions obtained from different calibrations and more technical problems need to be resolved. Just as MIRT can be considered either as a special case of factor analysis or an extension of unidimensional IRT, the multidimensional scale linking can be realized either by borrowing methods from factor analysis (Hirsch, 1988; Li, 1997; Min, 2003) or by extending the unidimensional IRT linking methods to the multidimensional situations (Oshima et al., 2000).

Hirsch's Method

Hirsch (1988) is possibly the first author to explore the feasibility and effectiveness of multidimensional linking and equating by using the common-examinee design. Hirsch presented three technical issues in multidimensional linking and provided three possible resolutions. The first issue is to establish scale transformations to keep the M2PL function invariant. The following transformation equations can be used for a two-dimension (dimension 1 and 2) M2PL model:

$$\theta_{i1}^{*} = \frac{\theta_{i1} - \mu_{1\theta}}{\sigma_{1\theta}}, \quad \theta_{i2}^{*} = \frac{\theta_{i2} - \mu_{2\theta}}{\sigma_{2\theta}}, \tag{1-43}$$

$$a_{j1}^{*} = \sigma_{1\theta} a_{j1}, \quad a_{j2}^{*} = \sigma_{2\theta} a_{j2}, \tag{1-44}$$

$$d_{j}^{*} = d_{j} + a_{j1}\mu_{1\theta} + a_{j2}\mu_{2\theta}, \tag{1-45}$$

where parameters with superscript "$*$" are transformed parameters on a new scale. The M2PL function is invariant by this transformation:

$$\begin{aligned}
&P(x_{ij} = 1 | \boldsymbol{\theta}_i^{*}; \mathbf{a}_j^{*}, d_j^{*}) \\
&= \frac{1}{1 + e^{-\left[\sigma_{1\theta} a_{j1}\left(\frac{\theta_{i1} - \mu_{1\theta}}{\sigma_{1\theta}}\right) + \sigma_{2\theta} a_{j2}\left(\frac{\theta_{i2} - \mu_{2\theta}}{\sigma_{2\theta}}\right) + (d_j + a_{j1}\mu_{1\theta} + a_{j2}\mu_{2\theta})\right]}} \\
&= \frac{1}{1 + e^{-\left[a_{j1}(\theta_{i1} - \mu_{1\theta}) + a_{j2}(\theta_{i2} - \mu_{2\theta}) + (d_j + a_{j1}\mu_{1\theta} + a_{j2}\mu_{2\theta})\right]}} \\
&= \frac{1}{1 + e^{-\left[a_{j1}\theta_{i1} + a_{j2}\theta_{i2} + d_j\right]}} \\
&= P\left(x_{ij} = 1 | \boldsymbol{\theta}_i; \mathbf{a}_j, d_j\right).
\end{aligned}$$

This scale transformation method can be extended to M2PL models with more than two dimensions. Hirsch's multidimensional scale linking method was based on the invariance of multidimensional function under the above transformations of item and ability parameters.

The second technical issue is that the correlation between dimensions obtained from the first calibration may be somewhat different from the correlation estimated from the second calibration due to some non-parallel items for common-examinee design. If this occurs, the parameter estimates from two calibrations are composites or linear combinations of different basis vectors. Therefore, it is necessary to transform the basis vectors from one calibration to those of the second calibration. This can be realized by transforming the two sets of ability parameter estimates of the common examinees from two calibrations so that they are as similar as possible.

The third technical issue is the joint rotational indeterminacy of the item discrimination and ability parameters. That is, the dimensions can be rotated and produce many possible sets of $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$ parameter estimates without affecting the M2PL item characteristic function. As suggested by Wang (1985), the procrustean rotation in factor analysis (Schonemann ,1966) can be used to transform the parameter estimates from one calibration to those from the other calibration.

Hirsch's linking method for the common-examinee design includes four steps. In the first step, two sets of item and ability parameters $(\hat{\boldsymbol{\theta}}_{F_i}, \hat{\mathbf{a}}_{F_j})$ and $\left(\hat{\boldsymbol{\theta}}_{E_i}, \hat{\mathbf{a}}_{E_j}\right)$ for the common examinees but on different metrics are estimated from two independent calibrations. In $(\hat{\boldsymbol{\theta}}_{F_i}, \hat{\mathbf{a}}_{F_j})$, $\hat{\boldsymbol{\theta}}_{F_i}$ is a $N \times m$ matrix, where N is the number of examines, and $\hat{\mathbf{a}}_{F_j}$ is a $n \times m$ matrix, where n is the number of items. In the second step, three transformations are used to obtain common basis vectors for the two sets of parameter estimates. The first transformation by $\mathbf{T}_1$ refers the discrimination parameter estimates from the first calibration ($\hat{\mathbf{a}}_{F_j}$) to a set of orthogonal basis vectors instead of the basis vectors defined by the ability estimates ($\hat{\boldsymbol{\theta}}_{F_i}$). The second transformation by $\mathbf{T}_2^{-1}$ refers the discrimination parameter estimates from the second calibration ($\hat{\mathbf{a}}_{E_j}$) to a set of orthogonal basis vectors instead of the basis vectors defined by the ability estimates ($\hat{\boldsymbol{\theta}}_{E_i}$). The third transformation by $\mathbf{T}_3 = \mathbf{T}_1 * \mathbf{T}_2^{-1}$ refers the discrimination parameter estimates from the first calibration ($\hat{\mathbf{a}}_{F_j}$) to a set of common basis vectors for both calibrations. In the third step, orthogonal procrustean transformation is used to rotate the ability estimates from the first calibration ($\hat{\boldsymbol{\theta}}_{F_i}$) to those from the second calibration ($\hat{\boldsymbol{\theta}}_{E_i}$). This fourth transformation matrix $\mathbf{T}_4$ can be found by minimizing the sum of squared difference between

each element of the two sets of ability parameter estimates ($\hat{\boldsymbol{\theta}}_{F_i}$) and ($\hat{\boldsymbol{\theta}}_{E_i}$). The method was called orthogonal procrustean transformation developed by Schonemann (1966). Specifically, suppose $\mathbf{S} = \hat{\boldsymbol{\theta}}_{F_i}^{'}\hat{\boldsymbol{\theta}}_{E_i}$, $\mathbf{SS}^{'} = \mathbf{PDP}^{'}$, and $\mathbf{S}^{'}\mathbf{S} = \mathbf{QDQ}^{'}$, then $\mathbf{T}_4 = \mathbf{PQ}^{'}$. Given the above four transformations, the means and standard deviations of the ability parameters for the common examinees from the two calibrations are estimated in the fourth step. For the common-examinee design, the linking parameters can be estimated by equating the means and standard deviations of the ability estimates from the first calibration ($\hat{\boldsymbol{\theta}}_{F_i}$) and those transformed from the second calibrations ($\hat{\boldsymbol{\theta}}_{F_i}^{*}$). The linking parameter estimates are then used to transform the parameter estimates which have already been transformed by the procedure described in the second step. For example, suppose one uses the common examinee design and the M2PL model with two dimensions, the following relations exist:

$$\frac{\theta_{F_{1i}} - \mu_{F_{1\theta}}}{\sigma_{F_{1\theta}}} = \frac{\theta_{E_{1i}} - \mu_{E_{1\theta}}}{\sigma_{E_{1\theta}}},$$

$$\frac{\theta_{F_{2i}} - \mu_{F_{2\theta}}}{\sigma_{F_{2\theta}}} = \frac{\theta_{E_{2i}} - \mu_{E_{2\theta}}}{\sigma_{E_{2\theta}}}. \tag{1-46}$$

So

$$\theta_{F_{1i}} = \frac{\theta_{E_{1i}} - \left[\mu_{E_{1\theta}} - \left(\frac{\sigma_{E_{1\theta}}}{\sigma_{F_{1\theta}}}\right)\mu_{F_{1\theta}}\right]}{\frac{\sigma_{E_{1\theta}}}{\sigma_{F_{1\theta}}}},$$

$$\theta_{F_{2i}} = \frac{\theta_{E_{2i}} - \left[\mu_{E_{2\theta}} - \left(\frac{\sigma_{E_{2\theta}}}{\sigma_{F_{2\theta}}}\right)\mu_{F_{2\theta}}\right]}{\frac{\sigma_{E_{2\theta}}}{\sigma_{F_{2\theta}}}}. \tag{1-47}$$

Let

$$M_{1\theta} = \mu_{E_{1\theta}} - \left(\frac{\sigma_{E_{1\theta}}}{\sigma_{F_{1\theta}}}\right)\mu_{F_{1\theta}},$$

$$M_{2\theta} = \mu_{E_{2\theta}} - \left(\frac{\sigma_{E_{2\theta}}}{\sigma_{F_{2\theta}}}\right)\mu_{F_{2\theta}}, \tag{1-48}$$

$$S_{1\theta} = \frac{\sigma_{E_{1\theta}}}{\sigma_{F_{1\theta}}},$$

$$S_{2\theta} = \frac{\sigma_{E_{2\theta}}}{\sigma_{F_{2\theta}}}. \tag{1-49}$$

Then the transformed parameters from E scale to F scale are

$$\theta^*_{F_{1i}} = \frac{\theta^*_{E_{1i}} - M_{1\theta}}{S_{1\theta}},$$

$$\theta^*_{F_{2i}} = \frac{\theta^*_{E_{2i}} - M_{2\theta}}{S_{2\theta}}, \tag{1-50}$$

$$a^*_{F_{1j}} = S_{1\theta} a^*_{E_{1j}},$$

$$a^*_{F_{2j}} = S_{2\theta} a^*_{E_{2j}}, \tag{1-51}$$

$$d^*_{F_j} = d^*_{E_j} + a^*_{E_{1j}} M_{1\theta} + a^*_{E_{2j}} M_{2\theta}, \tag{1-52}$$

where the parameters with "$*$" as superscript and "F" as subscript on the left side of equations are the final transformed parameters on F scale, and the parameters with "$*$" as superscript and "E" as subscript on the right side of equations are the transformed parameters on E scale by the first three transformations.

The function of this four-step scale linking procedure for M2PL model was evaluated by test equating results performed on both simulated and real data sets using the common-examinee

design (Hirsch, 1989). The equating results were examined by comparing the mean differences and the mean absolute differences of the true scores and ability estimates between the base tests and equated tests. Satisfactory equating was found for true scores but not for ability estimates.

Hirsch's linking method was originally developed for the common-examinee design. However, it can easily be modified to conduct scale linking for common-item nonequivalent groups design which is most usually used in test equating and DIF study. As Hirsch (1988) suggested, the basis vector transformation would be the same. The procrustean transformation would use the common item discrimination parameters instead of the ability parameters. The item difficulty parameter for each item would need to be regressed onto each of the ability dimension parameters and therefore produce one unique difficulty parameter for each of the dimensions (Reckse, 1985). Then the mean and sigma method would be used for the common item difficulty parameters for the final transformation. However, more study is needed to verify the adequacy of this modified linking procedure.

Li's Method

Compared with Hirsch's procedure, Li's (1997) multidimensional linking methods are more straightforward and consistent with MIRT computer estimation programs. Most MIRT programs solve the identification problem by requiring multidimensional abilities be distributed as multivariate normal MVN (**0, 1**). Therefore, the metric of the item parameter estimates is typically referred to orthogonal reference axes with unit length. Given this condition, one reference system can be transformed onto the other reference system by a composite transformation: an orthogonal procrustean transformation for re-rotating the reference system, a translation transformation for shifting the point of origin, and a single dilation for re-scaling unit length. Specifically the following equations are used in the reference system transformation:

$$\mathbf{a}^*_{F_j} = k\mathbf{T}'\mathbf{a}_{E_j}, \tag{1-53}$$

$$d^*_{F_j} = d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}, \tag{1-54}$$

$$\theta^*_{F_i} = (1/k)\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right). \tag{1-55}$$

It can be shown that the M2PL function is invariant to these transformations:

$$\begin{aligned}
&P(x_{ij} = 1 \mid \boldsymbol{\theta}^*_{F_i}; \mathbf{a}^*_{F_j}, d^*_{F_j}) \\
&= \frac{1}{1+e^{-\left\{\left[k\mathbf{T}'\mathbf{a}_{E_j}\right]\left[(1/k)\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right)\right] + \left[d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1+e^{-\left\{\left[k\mathbf{a}'_{E_j}\mathbf{T}\right]\left[(1/k)\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right)\right] + \left[d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1+e^{-\left\{\left[\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i} - \mathbf{a}'_{E_j}\mathbf{T}\mathbf{m}\right] + \left[d_{E_j} + \mathbf{a}'_{E_j}\mathbf{T}\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1+e^{-\left\{\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i} + d_{E_j}\right\}}} \\
&= P(x_{ij} = 1 \mid \boldsymbol{\theta}_{E_i}; \mathbf{a}_{E_j}, d_{E_j}).
\end{aligned}$$

The question is how to find **T**, **m**, and *k*. Li (1997) proposed several methods to estimate the scale linking parameters. The rotation matrix **T** can be estimated by orthogonal procrustean transformation procedure as mentioned in Hirsch's method above. Let $\mathbf{S} = \hat{\mathbf{a}}'_{F_j}\hat{\mathbf{a}}_{E_j}$, $\mathbf{SS}' = \mathbf{PDP}'$ and $\mathbf{S}'\mathbf{S} = \mathbf{QDQ}'$, then $\mathbf{T} = \mathbf{PQ}'$.

The origin shift coefficient **m** and unit change coefficient *k* can be estimated simultaneously by minimizing the sum of squared difference between test characteristic functions for the common items obtained from the two calibrations, which was originally developed by Stocking and Lord (1983) for the unidimensional linking:

$$f(\mathbf{m}, k) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j=1}^{n} P_{F_j}\left(\boldsymbol{\theta}; \hat{\mathbf{a}}_{F_j}, \hat{d}_{F_j}\right) - \sum_{j=1}^{n} P^*_{F_j}\left(\boldsymbol{\theta}; \hat{\mathbf{a}}^*_{F_j}, \hat{d}^*_{F_j}\right)\right]^2, \tag{1-56}$$

where *N* is the number of grid points of θ values.

The origin shift and unit change coefficients can also be estimated separately by different procedures. For example, the origin shift coefficient can be estimated by minimizing the sum of squared difference between the two difficulty parameter estimates obtained from two calibrations:

$$f(\mathbf{m}) = \sum_{j=1}^{n} \left(\hat{d}_{F_j} - \hat{d}^*_{F_j} \right)^2 , \tag{1-57}$$

where n is the number of common items. This was called least squares procedure (Li, 1997). The unit change coefficient can be estimated as the ratio of the square root of the maximum eigenvalues of the matrices $\hat{\mathbf{a}}'_{F_j} \hat{\mathbf{a}}_{F_j}$ and $\hat{\mathbf{a}}'_{E_j} \hat{\mathbf{a}}_{E_j}$ obtained from the two calibrations:

$$k = \frac{Maximum\left[sig\left(\hat{\mathbf{a}}'_{F_j} \hat{\mathbf{a}}_{F_j} \right) \right]}{Maximum\left[sig\left(\hat{\mathbf{a}}'_{E_j} \hat{\mathbf{a}}_{E_j} \right) \right]}, \tag{1-58}$$

where $sig(\cdot)$ represents the singular value or the nonnegative square roots of the eigenvalue. This was called ratio of eigenvalues procedure (Li, 1997). Similar to the least squares procedure for the estimation of origin shift coefficient, the unit change coefficient can also be estimated by minimizing the sum of squared difference between the two sets of discrimination parameters estimated from two calibrations. This is also referred as least squares procedure (Li, 1997):

$$f(k) = \sum_{j=1}^{n} \left(\hat{\mathbf{a}}_{F_j} - \hat{\mathbf{a}}^*_{F_j} \right)^2 . \tag{1-59}$$

The rotation matrix **T** and unit change coefficient k can also be estimated simultaneously by a least squares method developed for fitting one matrix to another through a rotation matrix, a translation vector, and a central dilation vector (Schonemann & Carroll, 1970). In this case, the rotation matrix and dilation scalar were estimated by minimizing the sum of squared errors of the following residual matrix:

$$\mathbf{E} = \left(k\hat{\mathbf{a}}'_{F_j}\mathbf{T}\right) - \hat{\mathbf{a}}'_{E_j}. \tag{1-60}$$

It can be shown that

$$k = \frac{trace\left(\mathbf{T}'\mathbf{a}_{CF_j}\right)\mathbf{a}'_{CE_j}}{trace\left(\mathbf{a}_{CF_j}\mathbf{a}'_{CF_j}\right)}, \tag{1-61}$$

where

$$\mathbf{a}_{CF_j} = \hat{\mathbf{a}}_{F_j} - \bar{\mathbf{a}}_{F_j}, \mathbf{a}_{CE_j} = \hat{\mathbf{a}}_{E_j} - \bar{\mathbf{a}}_{E_j}, \tag{1-62}$$

with $\bar{\mathbf{a}}_{F_j}$ as the mean of $\hat{\mathbf{a}}_{F_j}$ and $\bar{\mathbf{a}}_{E_j}$ as the mean of $\hat{\mathbf{a}}_{E_j}$. This was called ratio of trace procedure (Li, 1997). The translation vector was not estimated by this method because item discriminations can not provide information about origin shift.

Comparing the effect of different combinations of reference, translation, and dilation transformation procedures on the multidimensional linking parameters estimation, Li (1997) found that the most appropriate MIRT linking method is the combination of procrustean rotation approach (for dimensional transformation), the ratio of trace procedure (for dilation), and the least square procedure (for translation). This linking method could produce accurate estimation of item parameters, approximately equivalent estimation of ability parameters, but unsatisfactory true score estimation.

Min's Method

Min (2003) challenged Li's (1997) two reasons for using a single dilation parameter, that is, mathematical tractability and the assumption of constant variance across dimensions, and argued that one single dilation is insufficient for describing the scale unit changes for multiple dimensions. Two independent calibrations may change the scales of the multidimensional dimensions to different degrees. To address this problem, Min (2003) modified Li's (1997) method by replacing the single dilation parameter with a diagonal dilation matrix to model

different unit changes on different dimensions. The reference system transformations are performed as follows:

$$\mathbf{a}^*_{F_j} = \mathbf{K}'\mathbf{T}'\mathbf{a}_{E_j}\,, \tag{1-63}$$

$$d^*_{F_j} = d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}\,, \tag{1-64}$$

$$\theta^*_{F_i} = \mathbf{K}^{-1}\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right), \tag{1-65}$$

where **K** is a diagonal dilation matrix. It can be shown that the M2PL function is invariant to these transformations:

$$\begin{aligned}
&P(x_{ij} = 1 \mid \boldsymbol{\theta}^*_{F_i}; \mathbf{a}^*_{F_j}, d^*_{F_j}) \\
&= \frac{1}{1 + e^{-\left\{\left[\mathbf{K}'\mathbf{T}'\mathbf{a}_{E_j}\right]\left[\left(\mathbf{K}^{-1}\right)\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right)\right] + \left[d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1 + e^{-\left\{\left[\mathbf{a}'_{E_j}\mathbf{T}\mathbf{K}\right]\left[\left(\mathbf{K}^{-1}\right)\left(\mathbf{T}^{-1}\boldsymbol{\theta}_{E_i} - \mathbf{m}\right)\right] + \left[d_{E_j} + \left(\mathbf{a}'_{E_j}\mathbf{T}\right)\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1 + e^{-\left\{\left[\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i} - \mathbf{a}'_{E_j}\mathbf{T}\mathbf{K}\mathbf{m}\right] + \left[d_{E_j} + \mathbf{a}'_{E_j}\mathbf{T}\mathbf{m}\right]\right\}}} \\
&= \frac{1}{1 + e^{-\left\{\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i} + d_{E_j}\right\}}} \\
&= P(x_{ij} = 1 \mid \boldsymbol{\theta}_{E_i}; \mathbf{a}_{E_j}, d_{E_j}).
\end{aligned}$$

For two-dimensional model, **K** becomes $\begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}$ where k_1 is the dilation parameter for the first dimension, and k_2 for the second dimension. The least square method (Schonemann & Carroll, 1970) of estimating a rotation matrix, a translation vector, and a central dilation vector for fitting two matrixes can be followed to find **T**, **K**, and **m** in the transformation equations (Min, 2003). Mathematically Li's (1997) method and Min's (2003) method produce the same solution for **T** and **m** and the only difference of linking results comes from the different dilation parameters.

Reckase and Martineau (2004) identified an important weakness in Li's (1997) and Min's (2003) method for MIRT models with high dimensionality and provided a solution to the problem by employing a non-orthogonal procrustean transformation. However, this approach needs to be examined by further empirical studies.

Oshima and Colleagues' Method

All multidimensional linking methods mentioned above borrowed an important procedure, procrustean rotation, from factor analysis to transform the dimensional axes. Oshima et al. (2000) extended four scale linking methods within IRT from unidimensional to multidimensional models. According to their methods, the following equations were used to transform the IRT parameters on one scale E to another scale F (to distinguish IRT linking methods from the factor analysis methods described above, different indices for linking parameters are used). For person i and item j,

$$\mathbf{a}^*_{F_j} = \left(\mathbf{A}^{-1}\right)' \mathbf{a}_{E_j}, \tag{1-66}$$

$$d^*_{F_j} = d_{E_j} - \mathbf{a}'_{E_j} \mathbf{A}^{-1} \boldsymbol{\beta}, \tag{1-67}$$

$$\boldsymbol{\theta}^*_{F_i} = \mathbf{A}\boldsymbol{\theta}_{E_i} + \boldsymbol{\beta}, \tag{1-68}$$

where the rotation matrix $\mathbf{A}_{\mathrm{m}\times\mathrm{m}}$ adjusts the variances and covariances of the ability dimensions (scale), and the translation vector $\boldsymbol{\beta}_{m\times 1}$ changes the means of the ability dimensions (location) on the two scales. The model indeterminacy can be shown as the following:

$$
\begin{aligned}
&P(x_{ij}=1|\boldsymbol{\theta}^*_{F_i};\mathbf{a}^*_{F_j},d^*_{F_j})\\
&=\frac{1}{1+e^{-\left\{\left[\left(\mathbf{A}^{-1}\right)'\mathbf{a}_{E_j}\right]'\left[\mathbf{A}\boldsymbol{\theta}_{E_i}+\boldsymbol{\beta}\right]+\left[d_{E_j}-\mathbf{a}'_{E_j}\mathbf{A}^{-1}\boldsymbol{\beta}\right]\right\}}}\\
&=\frac{1}{1+e^{-\left\{\left[\mathbf{a}'_{E_j}\mathbf{A}^{-1}\right]\left[\mathbf{A}\boldsymbol{\theta}_{E_i}+\boldsymbol{\beta}\right]+\left[d_{E_j}-\mathbf{a}'_{E_j}\mathbf{A}^{-1}\boldsymbol{\beta}\right]\right\}}}\\
&=\frac{1}{1+e^{-\left\{\left[\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i}+\mathbf{a}'_{E_j}\mathbf{A}^{-1}\boldsymbol{\beta}\right]+\left[d_{E_j}-\mathbf{a}'_{E_j}\mathbf{A}^{-1}\boldsymbol{\beta}\right]\right\}}}\\
&=\frac{1}{1+e^{-\left\{\mathbf{a}'_{E_j}\boldsymbol{\theta}_{E_i}+d_{E_j}\right\}}}\\
&=P(x_{ij}=1|\boldsymbol{\theta}_{E_i};\mathbf{a}_{E_j},d_{E_j}).
\end{aligned}
$$

As in the unidimensional IRT, suppose that two nonequivalent groups of examinees take common test items and independent calibrations produce two sets of parameter estimates ($\hat{\mathbf{a}}_{F_j},\hat{d}_{F_j}$) and ($\hat{\mathbf{a}}_{E_j},\hat{d}_{E_j}$). These two sets of parameter estimates are on different scales F and E, and scale linking needs to be conducted to place the two sets of parameter estimates on a common scale. Using the above equations, ($\hat{\mathbf{a}}_{E_j},\hat{d}_{E_j}$) on E scale can be transformed to the F scale ($\hat{\mathbf{a}}^*_{F_j},\hat{d}^*_{F_j}$). The values of the two sets of item parameter estimates ($\hat{\mathbf{a}}_{F_j},\hat{d}_{F_j}$) and ($\hat{\mathbf{a}}^*_{F_j},\hat{d}^*_{F_j}$) should be similar due to the invariance of common item characteristic in IRT. Unidimensional IRT linking methods can be extended to multidimensional IRT model to minimize some functions of the difference between the two sets of item parameters. Again, the question is how to find the values of **A** and **β** so that the connection between the two scales can be established**.**

The direct method. This method was a multivariate extension of the minimum chi-square linking method for unidimensional IRT model (Divgi, 1985). The values of **A** and **β** are estimated by minimizing the sum of squared difference between the two sets of item parameter estimates over all items. However, the direct method is different from the original method in that

it does not consider the variance-covariance matrix of sampling errors for item parameter estimates in the function:

$$f(\mathbf{A},\boldsymbol{\beta}) = \frac{1}{n(m+1)}\left\{\sum_{j=1}^{n}\sum_{k=1}^{m}[\hat{a}_{F_{jk}} - \hat{a}^{*}_{F_{jk}}]^2 + \sum_{j=1}^{n}[\hat{d}_{F_j} - \hat{d}^{*}_{F_j}]^2\right\}, \qquad (1\text{-}69)$$

where n is the number of items,

m is the number of ability dimensions, and

($\hat{a}^{*}_{F_{jk}}, \hat{d}^{*}_{F_j}$) are transformed parameter estimates from E scale to F scale.

The equated function method. This method is the multidimensional extension of the mean and sigma methods for the unidimensional IRT model (Loyd & Hoover, 1980; Marco, 1977). A more general system of scale linking equations is used to specify that some functions of the common item parameters from the first calibration ($\hat{\mathbf{a}}_{F_j}, \hat{d}_{F_j}$) are equal to the same functions of the transformed common item parameters from the second calibration ($\hat{\mathbf{a}}^{*}_{F_j}, \hat{d}^{*}_{F_j}$). The transformed item parameter estimates can be obtained by using the above scale transformation equations with the linking parameters **A** and **β.** The values of **A** and **β** are estimated by minimizing the sum of squared difference between the same functions of the two sets of selected elements of the estimated ($\mathbf{a}_{F_j}, d_{F_j}$) and ($\mathbf{a}^{*}_{F_j}, d^{*}_{F_j}$). The number of functions needed (p) depends on the number of dimensions (m) or elements in **A** and **β,** with $p = m^2 + m$. For example, in the two dimensional case ($m = 2$), four parameters in **A** and two parameters in **β** need to be estimated. Therefore, six functions are required to estimate the six linking parameters and they could be the means of $\hat{a}_{j1}$, $\hat{a}_{j2}$, and $\hat{d}_j$ for the first and second halves of the common items (or other block of items).

The scale linking functions are flexible in terms of which item parameter estimates to use and what function to use. Different systems of scale linking functions may produce different values of the linking parameters **A** and **β**. The quality or appropriateness of linking functions can be evaluated by their stability across random examinee samples, the character of the common item sets, and the true values of the linking parameters (Davey, et al., 1996). For example, if the mean is the chosen linking function, the function to be minimized is

$$f(\mathbf{A},\boldsymbol{\beta}) = \frac{1}{p}\sum_{j=1}^{p}\left(\hat{\mu}_{F_p} - \hat{\mu}^*_{F_p}\right)^2, \tag{1-70}$$

where $\hat{\mu}_{F_1}, \hat{\mu}_{F_2}, \cdots \hat{\mu}_{F_p}$ are the estimated means of p separate sets of elements of

the estimated ($\mathbf{a}_{F_j}, d_{F_j}$), and

$\hat{\mu}^*_{F_1}, \hat{\mu}^*_{F_2} \cdots \hat{\mu}^*_{F_p}$ are the estimated means of p separate sets of elements of the

estimated ($\mathbf{a}^*_{F_j}, d^*_{F_j}$).

The test characteristic function method. This method is an extension of the test response function method developed by Stocking and Lord (1983) for the unidimensional IRT model:

$$f(\mathbf{A},\boldsymbol{\beta}) = \frac{1}{q}\sum_{\theta} W_\theta \left[\sum_{j=1}^{n} P_{F_j}\left(\boldsymbol{\theta};\hat{\mathbf{a}}_{F_j}, \hat{d}_{F_j}\right) - \sum_{j=1}^{n} P^*_{E_j}\left(\boldsymbol{\theta};\hat{\mathbf{a}}^*_{F_j}, \hat{d}^*_{F_j}\right)\right]^2, \tag{1-71}$$

where q is the number of matching $\boldsymbol{\theta}$ vectors,

W_θ is the weight taken at different θ values.

The W_θ is used to emphasize that some θ values are more important than others to estimate the linking parameters. The weight can also be considered equal along the ability scale.

The item characteristic function method. This method is the multidimensional generalization of the item response function method for unidimensional IRT model (Haebara, 1980):

$$f(\mathbf{A},\boldsymbol{\beta}) = \frac{1}{n \times q} \sum_{\theta} W_{\theta} \sum_{j=1}^{n} \left[P_{F_j}\left(\boldsymbol{\theta}; \hat{\mathbf{a}}_{F_j}, \hat{d}_{F_j}\right) - P^*_{E_j}\left(\boldsymbol{\theta}; \hat{\mathbf{a}}^*_{F_j}, \hat{d}^*_{F_j}\right) \right]^2 . \qquad (1\text{-}72)$$

Based on a simulation study comparing the four IRT linking methods under different ability distributions (Oshima et al., 2000), all of the four methods were acceptable under almost any of the minimization criteria and offered dramatic improvement over not linking at all. It was also found that the test characteristic function method and item characteristic function method were more stable and recovered the true linking parameters better than the direct method and equated function method.

The multidimensional linking methods developed by Hirsch (1988), Li (1997), Oshima et al. (2000), and Min (2003) can all be directly or indirectly performed for the common-item nonequivalent groups design, which have been a widely used in test equating (Kolen & Brennan, 2004). Accordingly those methods have the potential for establishing calibrated item pool and exploring DIF. Another multidimensional linking method proposed by Thompson, Nering, and Davey (1997) can be used for test equating in a design without common items or examinees. With the assumption of the same origin, axes, and correlation between axes for the two randomly equivalent groups of examinees, this method solve the rotational indeterminacy by identifying similar item content clusters on different tests and then rotating them in the same multidimensional-reference system. Further studies need to be conducted to evaluate the performance of this method.

Multidimensional scale linking is a new research area. There have been very few studies conducted for each of the proposed methods (Hirsch, 1988; Li, 1997; Min, 2003; Oshima et al.,

2000) and even fewer studies for comparing different methods in the literature. Therefore, it is currently difficult to evaluate the function of different methods. The only comparison study by Min (2003) compared Li's method, Min's method, and Oshima and colleagues' test characteristic function method in terms of accuracy and stability of scale transformations under different conditions varying in sample size, structure of dimensions, and ability distribution. The results indicate that both Oshima and colleagues' and Min's methods were better in transforming discrimination parameters than Li's method, and Min's and Li's methods performed better than Oshima and colleagues' method in transforming the difficulty related parameters. In addition, Oshima and colleagues' method performed better than Min's and Li's methods in transforming test true scores, and Li's and Min's methods were better than Oshima and colleagues' method in maintaining the structure of dimensions through orthogonal rotation.

Purpose of the Study

Based on the literature review of the multidimensional linking methods, Li's methods have been evaluated under various circumstances such as different linking procedures, sample sizes, equating situations, number of anchor items, linking situations, and ability distributions (Li, 1997). Min's method has also been examined with comparison with other methods under different conditions including different sample sizes, dimensional structures, and ability distributions. The performance of Oshima and colleagues' four IRT linking methods has been examined under fewer testing conditions, that is, for different ability distributions, using simulation study with only 20 replications (Oshima et al., 2000). A comparison study (Min, 2003) indicates that one of the four IRT linking methods, that is, the test characteristic function method, outperformed other methods in transforming item discrimination parameter estimates and equating true score estimates. This suggests that the IRT procedures are promising methods for multidimensional linking and equating. Further studies are needed to examine the

performance of these four methods under more testing conditions. As an extension of previous research (Oshima et al., 2000), the purpose of this study was to evaluate the performance of the four multidimensional IRT scale linking methods, the direct method, equated function method, test characteristic function method, and item characteristic function method, under various testing conditions, which include different test structures, test lengths, sample sizes, and examinees' ability distributions.

CHAPTER 2
METHODOLOGY

A comprehensive review of the unidimensional scale linking and test equating (Cook & Petersen, 1987) provides us a framework for exploring the performance of multidimensional scale linking methods. According to Cook and Petersen's discussion, the results of linking and equating depend on linking or equating methods, sample characteristics, and properties of the common items. In addition, the multidimensional structure underlying the test item responses makes scale linking more complicated and should be considered as one important testing condition. In this simulation study, the performance of the four MIRT scale linking methods (Oshima et al., 2000) for the common-item nonequivalent groups design was evaluated with the compensator compensatory M2PL model under different testing conditions, including different test structures, test lengths, sample sizes, and examinees' ability distributions. The M2PL model had two dimensions.

Design

Independent Variables or Experimental Conditions

IRT linking method. This study was to evaluate the performance of the four multidimensional IRT scale linking methods proposed by Oshima et al. (2000): the direct method, equated function method, test characteristic function method, and item characteristic function method (see the section Multidimensional IRT Scale Linking in Chapter 1 for detailed description). The equated function, test characteristic function, and item characteristic function methods were implemented in a manner consistent with the implementation in Oshima and colleagues' study (2000). For the equated function method, the means of $\hat{a}_{j1}$, $\hat{a}_{j2}$, and $\hat{d}_j$ for the first and second halves of the items were used as the equated function. For the test and item characteristic function methods, seven equally spaced θ_1 points from -4 to 4 and seven equally

spaced θ_2 points from -4 to 4, making 7 x 7 = 49 grid points, were used with equal unit weight along the ability scale. The four IRT linking methods have been compared under different ability distributions (Oshima et al., 2000). It is unknown how they perform under other circumstances. Therefore, this study can be considered as an extension of Oshima and colleagues' study (2000) from one testing condition (ability distribution) to various testing conditions (see the following for the detail).

Test structure. In IRT, the test dimensionality for a particular population is the minimum number of latent abilities required to produce a monotone and locally independent model (McDonald, 1981, 1997; Stout, 1990). In the geometrical representation of a test structure, the coordinate axes of a multidimensional space is defined by a complete set of latent abilities examined by the test, and each item is described by a vector in the space with its orientation representing the ability composite that is best measured by the item (Ackerman, 1994, 1996; Reckase, 1985, 1991). According to the literature review by Tate (2003), based on the number and nature of the abilities required for the response to each item in the test, there are three types of test structure: simple structure, approximate simple structure, and complex structure. In the simple structure, all item vectors are exactly aligned with one of the axes in the multidimensional space after an appropriate rotation, so all the items under each dimension measure the same ability. If all item vectors are approximately aligned with one of the multiple axes and therefore the contribution of one ability is dominant over the contribution of all other abilities, the test has approximate simple structure. In complex test structure, the response to one item depends on more than one ability. The first type of structure has been considered as an ideal one and the second and third types as more realistic item structures (Kim, 1994; Roussos, Stout, & Marden, 1998). Following the method in previous studies (Batley & Boss, 1993; Min, 2003; Mroch &

Bolt, 2006; Oshima et al., 1997; Oshima & Miller, 1992; Tate, 2003), two types of two-dimension test structure were created by using the three MIRT item characteristics: MDISC, MDIFF, and direction (Ackerman, 1994; Reckase, 1985; Reckase, 1997a; Reckase & McKinley, 1991). In the approximate simple structure, there were two sets of items: The responses to the first half items depended on one composite ability with the first dimension as the dominant dimension and the second dimension as the minor dimension; The responses to the second half items depended on another composite ability with the second dimension as the dominant dimension and the first dimension as the minor dimension. In the complex test structure, there were four sets of items with equal number of items in each of the set. Two sets of items loaded heavily on one of the two dimensions and lightly on the other dimension, and the remaining two sets loaded heavily on both dimensions.

Test length. The test items are used to establish the common metric for the two sets of parameter estimates obtained in separate calibrations. Therefore, the feature of items is very important for sale linking. The estimation of linking parameters depends not only on the number of items, but also on the characteristics of the item parameters. Based on some literature reviews (Brennan 1987; Cook & Petersen 1987; Kolen & Brennan, 2004), 15-30 common items are necessary for unidimensional IRT linking, although the required number also depends on other conditions, such as the linking methods, examinees' ability distributions, and characteristics of the items. Different numbers of items have been used in multidimensional linking studies. Li (1997) used 15 and 25 items in his study and found that the number of items had a significant influence on the stability of transformation parameter estimates for multidimensional linking. Oshima et al., (2000) created 40 item parameters to examine the performance of their four IRT multidimensional linking procedures. Twenty items were used in Min's study (2003) to compare

three multidimensional linking methods. In this study, 20 and 40 items were used to evaluate the four MIRT linking methods under different numbers of items with the consideration that more items may be needed for MIRT than unidimensional IRT linking.

Sample size. Theoretically the performance of linking methods depends on the accuracy of parameter estimates and parameter estimation is affected by the sample size (Li, 1997). So the linking function depends in some extent on different sample sizes. Compared with unidimensional IRT models, a large number of examinees are required for MIRT calibration because more parameters need to be estimated. Based on some MIRT researchers' (Ackerman, 1994; Carlson, 1987) recommendations, 2000 examinees is a reasonable sample size to obtain satisfactory item parameter estimates for compensatory multidimensional model. Reckase (1997a) reported that NORHARM (Fraser & McDonald, 1988) and TESTFACT (Wilson, Wood, & Gibbons, 1987) generally produced stable parameter estimates for long tests and sample sizes exceeding 1000 cases. A comprehensive study (Tate, 2003) using both simulated and real data found that most of the often-used multidimensional computer programs performed well for the sample size of 2000 examinees. To acquire stable item parameter estimates, Hirsch (1988) used 2000 examinees to evaluate the proposed multidimensional equating. In his first study, Li (1997) used three different sample sizes, 1000, 2000, and 4000, to examine the performance of three multidimensional linking methods and found that the sample size had a prominent role in estimating transformation parameters. Li used 2000 examinees to evaluate the best linking method in his second study (Li, 1997). Min (2003) also found the significant effect of sample sizes, 500, 1000, and 2000 on the accuracy and stability of different multidimensional linking methods and suggested that the sample of 500 examinees showed unreliable results and the sample of 1000 showed somewhat acceptable outcomes (note that approximate simple structure

and complex structure were used in the study). It is not unusual in testing practice that the sample size is less than 1000 especially in non-achievement area and the performance of the four IRT linking methods need to be evaluated under this condition. In this study, three different sample sizes, 500, 1000, 2000, were used to examine the robustness of the four multidimensional IRT scale linking methods against parameter estimation errors. As defined in other studies (Li, 1997; Min, 2003), the sample size of 2000 is the base for comparing the effect of different parameter estimation errors. The sample size of 500 can be used to examine the robustness of IRT scale linking for small sample size. The sample size of 1000 was used to examine the effect of sample size between 500 and 2000 on scale linking and it was also consistent with a study using multidimensional linking for identifying differential item functioning (Oshima et al., 1997).

Examinees' ability distribution. Based on the review by Kolen and Brennan (2004), the performance of scale linking also depends on the similarity between the two groups of examinees. The more similar the groups are, the more adequate the linking will be. Large difference between groups may produce significant problems in estimating scale linking parameters. Groups of examinees may differ in many characteristics, such as cultural background, attitude, motivation, and personality. A comprehensive review on population invariance in equating and linking (Kolen, 2004) found that equating is population dependent except under highly restrictive conditions, such as two test forms with similar content, difficulty, and reliability. This suggests that scale linking parameters that are used to obtain the equivalent scores should also be dependent on the populations used in the estimation. The ability distribution is an important characteristic of the examinees and has a significant influence on test equating and scale linking under both unidimensional (Cook et al., 1985) and multidimensional circumstances (Li, 1997; Min, 2003). As summarized by Cook and Petersen (1987), the

similarity of ability distribution between groups also affects other conditions required for test equating, such as the number of common items.

Groups of examinees may differ from each other in terms of mean, variance, and covariance of the dimensions. Oshima et al. (2000) examined the four multidimensional IRT scale linking methods under six conditions of the ability distributions across two groups: no difference at all; differences in θ variances; differences in θ correlations; differences in θ means; differences in θ means and variances; differences in θ means and correlations. Min (2003) used four conditions similar to those investigated by Oshima et al, such as differences in θ correlations; differences in θ correlations and means; and differences in θ correlations, means, and variances. However, in all conditions the ability dimensions were uncorrelated in the base group. In education and psychology, most constructs and dimensions within a construct are correlated. Two groups should have similar structure of construct before scale linking and equating are conducted. Given these two considerations and to keep the scope of the study manageable, correlations between dimensions were set at the same level, but not zero, across all groups and the two groups varied only in ability level and variance. One purpose of this study was to explore two-dimensional linking methods under the following four ability distributions: no difference at all, differences in θ means, differences in θ variances, and differences in θ means and variances (See Table 2-1 for the detail).

Dependent Variables or Evaluation Criteria

Different statistics have been used to evaluate multidimensional linking methods. Bias and root mean square error (RMSE) are often used to evaluate the accuracy and stability of results across replications of experiment in IRT simulation studies. For example, using a common examinee design, Hirsch (1988) evaluated the effectiveness of multidimensional linking

and equating by examining the means and standard deviations of the differences and absolute differences between the true scores, ability estimates, test characteristic response surfaces, and contour plots of the common examinees on the base and equated tests. Li (1997) used bias and RMSE to evaluate three multidimensional linking methods, but he used both the bias and RMSE of linking parameters and item and ability parameters over replications in his study. Oshima et al. (2000) compared the means, standard deviations, bias, and RMSE of linking parameters for different methods.

Another criterion for common item scale linking in IRT framework is to evaluate how small the differences are between the item parameter estimates for base group and the transformed item parameter estimates for equated group across the common items (Min, 2003; Min & Kim, 2003). This criterion was used in this study. Specifically, the common item nonequivalent groups design was used and simulation was performed to create the data for both base and equated groups. The parameters for the two groups were estimated and then transformed onto a common scale. Specifically, the parameter estimates for equated groups were transformed onto the scale for the base groups by using the transformation equations described before. The linking coefficients in the transformation equations, $\mathbf{A}$ and $\boldsymbol{\beta}$, were estimated through the four IRT multidimensional linking methods. After the common item parameters estimated from base and equated groups were placed on the same scale, the performance of the four linking methods were evaluated by examining the differences between the two sets of item parameter estimates. The mean difference and difference variation across replications (*r)* for each item were used to evaluate the accuracy and stability of the four linking methods, as described by the following statistics:

$$M_{diff}\left(\mathbf{a}_j\right)=\frac{\sum_{r=1}^{r} diff}{r}, \tag{2-1}$$

$$SD_{diff}\left(\mathbf{a}_j\right)=\sqrt{\frac{\left(diff-\overline{diff}\right)^2}{r}}, \tag{2-2}$$

where

$$diff=\hat{\mathbf{a}}^*_{F_j}-\hat{\mathbf{a}}_{F_j}, \tag{2-3}$$

$$M_{diff}\left(d_j\right)=\frac{\sum_{r=1}^{r} diff}{r}, \tag{2-4}$$

$$SD\left(d_j\right)=\sqrt{\frac{\left(diff-\overline{diff}\right)^2}{r}}, \tag{2-5}$$

where

$$diff=\hat{d}^*_{F_j}-\hat{d}_{F_j}. \tag{2-6}$$

Procedure

Data Generation

The following compensatory, two-parameter, two-dimension IRT model was used to create the item responses with different testing conditions described above:

$$P(x_{ij}=1|\theta_i;a_j,d_j)=\frac{1}{1+e^{-D\left(a_{j1}\theta_{i1}+a_{j2}\theta_{i2}+d_j\right)}}. \tag{2-7}$$

First, five sets of ability parameters for each of the three sample sizes (500, 1000, and 2000) with multivariate normal distributions with various means, variances, and covariances were generated. One set of ability parameters was used for the base group and the other for the four equated groups (see Table 2-1 for the five group ability distributions).

Second, two sets of item parameters (one with 20 items and another with 40 items) for each of the two test structures (approximate simple structure and complex structure) were created using the three MIRT item characteristics: MDISC, MDIFF, and direction (Ackerman, 1994; Reckase, 1985; Reckase, 1997a; Reckase & McKinley, 1991). Based on the pooled results from past empirical studies (Reckase, 1985; Reckase, 1997a; Reckase & McKinley, 1991), the estimated MIDSC has a lognormal distribution with mean of 1.37 and standard deviation of 0.54 and the estimated MDIFF has a normal distribution with mean of 0.28 and standard deviation of 0.69. The item parameters of MDISC and MDIFF in this study were selected randomly from lognormal and normal distributions with the same value of means and standard deviations. The test structure was created by manipulating the angle of each item with the first dimension. For the items that loaded on one dominant dimension, the angle between the item and its dominant dimension was selected from a lognormal distribution with mean of 10° and standard deviation of 2°. For the items loaded heavily on both dimensions, the angle between each item and two dimensions were selected from a normal distribution with mean of 45° and standard deviation of 10°. Next, the discrimination parameters, a_1, a_2, and the difficulty parameter, d, were computed by the following formula ((Ackerman, 1994; Reckase, 1985; Reckase, 1997a; Reckase & McKinley, 1991):

$$a_1 = MDISC * \cos\alpha_1, \tag{2-8}$$

$$a_2 = MDISC * \cos\alpha_2, \tag{2-9}$$

$$d = -MDISC * MDIFF\ . \tag{2-10}$$

(See Table 2-2, 2-3, 2-4, and 2-5 for specific parameter values for different test structures with different test lengths)

Next, dichotomous item responses were created using the two-parameter and two-dimension IRT model described by Equation 2-7.

To produce more precise and stable results, replications were conducted for each of the combinations of testing conditions. In IRT simulation studies, the number of replications depends on the purpose of the study, the desire of minimizing the sampling variance of the estimated parameters, and the need for statistical tests of results (Harwell, Stone, Hsu, & Kirisci, 1996). The previous studies on multidimensional linking or equating methods used 0 (Hirsch, 1988), 20 (Oshima et al., 2000), 50 (Min, 2003), 100, and 200 (Li, 1997) replications to evaluate the accuracy and stability of linking or equating results. Based on Harwell and colleagues' (1996) recommendation of using a minimum of 25 replications for IRT simulation studies and given the level of complexity of this study, 500 replications were used for each of the combinations of testing conditions to evaluate the accuracy and stability of the four multidimensional IRT linking methods.

Parameter Estimation

The parameters of MIRT models can be estimated using different methods and computer programs. The often used estimation methods include unweighted least squares (ULS) factor analysis of tetrochoric correlations, weighted least squares (WLS) analysis of the matrix of polychoric correlations, and robust WLS analysis methods performed by MPLUS (Muthen & Muthen, 1998), least squares estimation method based on the matrix of raw product moments of item scores by NOHARM (Fraser & McDonald, 1988), marginal maximum likelihood estimation method by TESTFACT (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 1999). The study focusing on model parameters recovery by Knol and Berger (1991) suggests that "for multidimensional data a common factor analysis on the matrix of tetrachoric correlations performs at least as well as the theoretically appropriate multidimensional item response models"

(p. 457). A study comparing TESTFACT and NOHARM (Gosz & Walker, 2002) found that NOHARM provided better solutions for predicting item performance. The comprehensive comparison study by Tate (2003) found that MPLUS, NOHARM, and TESTFACT performed reasonably well over a relatively wide range of conditions in assessing the test structure and estimating parameters. This result was confirmed by another recent study (Stone & Yeh, 2006). Based on these studies, all these methods can provide satisfactory estimation for model parameters. NOHARM was used in this study due to its consistently good performance in previous studies.

After the MIRT item parameters were estimated by NOHARM, the linking parameters estimated by the four multidimensional IRT linking methods (direct method, equated function method, test characteristic function method, and item characteristic function method) were obtained by the computer program IPLINK, which was developed by Lee and Oshima (1996).

Result Analysis

Some previous multidimensional linking studies used descriptive analysis (Hirsch, 1988; Oshima et al., 2000) and some studies used both descriptive and inferential analysis (Li, 2000; Min, 2003). In this study, the means and standard deviations of differences between the item parameter estimates for base group ($\hat{\mathbf{a}}_{F_j}, \hat{d}_{F_j}$) and the transformed item parameter estimates for equated group ($\hat{\mathbf{a}}^*_{F_{jk}}, \hat{d}^*_{F_j}$) across 500 replications were compared under different testing conditions. Specifically, the accuracy and stability of the four multidimensional IRT linking methods were evaluated by examining the mean differences and difference variations of a_1, a_2, and d for all items in the test under different testing conditions.

Based on the experimental conditions described above, there are 5 factors in this study: multidimensional linking method (4), test structure (2), test length (2), sample size (3), and

ability distribution (4). Therefore, the total number of experimental conditions is $4 \times 2 \times 2 \times 3 \times 4 =$ 192. Five hundred replications were conducted for each of the conditions.

Table 2-1. Ability distributions for examinee groups

Base Group		Group1		Group2		Group3		Group4	
μ	Σ	μ	Σ	μ	Σ	μ	Σ	μ	Σ
$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$	$\begin{bmatrix} .5 \\ .5 \end{bmatrix}$	$\begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} .8 & .4 \\ .4 & .8 \end{bmatrix}$	$\begin{bmatrix} .5 \\ .5 \end{bmatrix}$	$\begin{bmatrix} .8 & .4 \\ .4 & .8 \end{bmatrix}$

Note: All the correlations between dimensions are .5.

Table 2-2. Item parameters for 20 items with approximate simple structure

Item	a_1	a_2	d	MDISC	MDIFF	α_1	α_2
1	1.12	0.18	-0.70	1.13	0.62	9	81
2	2.23	0.52	-0.23	2.29	0.10	13	77
3	1.39	0.24	-2.19	1.41	1.55	10	80
4	1.02	0.16	-0.75	1.03	0.73	9	81
5	1.68	0.33	-1.18	1.71	0.69	11	79
6	0.98	0.15	0.77	0.99	-0.78	9	81
7	1.24	0.22	1.12	1.26	-0.89	10	80
8	0.94	0.13	-1.26	0.95	1.33	8	82
9	1.65	0.32	-1.86	1.68	1.11	11	79
10	2.01	0.46	-1.46	2.06	0.71	13	77
11	0.30	1.30	-0.72	1.33	0.54	77	13
12	0.17	1.09	0.17	1.10	-0.15	81	9
13	0.33	1.86	0.51	1.89	-0.27	80	10
14	0.08	0.63	-0.07	0.63	0.11	83	7
15	0.14	0.99	-1.38	1.00	1.38	82	8
16	0.17	1.09	-2.66	1.10	2.42	81	9
17	0.20	1.15	-0.48	1.17	0.41	80	10
18	0.13	0.84	-0.68	0.85	0.80	81	9
19	0.38	2.37	-0.77	2.40	0.32	81	9
20	0.22	1.04	-0.73	1.06	0.69	78	12

Table 2-3. Item parameters for 40 items with approximate simple structure

Item	a_1	a_2	d	MDISC	MDIFF	α_1	α_2
1	2.29	0.44	1.28	2.33	-0.55	11	79
2	1.10	0.21	0.21	1.12	-0.19	11	79
3	1.44	0.20	-1.45	1.45	1.00	8	82
4	0.57	0.09	0.28	0.58	-0.48	9	81
5	0.92	0.16	-1.59	0.93	1.71	10	80
6	0.96	0.20	0.45	0.98	-0.46	12	78
7	1.18	0.23	-0.79	1.20	0.66	11	79
8	1.12	0.16	-1.07	1.13	0.95	8	82
9	0.98	0.19	0.80	1.00	-0.80	11	79
10	1.90	0.30	-0.94	1.92	0.49	9	81
11	0.55	0.14	0.27	0.57	-0.48	14	76
12	1.35	0.26	0.40	1.38	-0.29	11	79
13	1.16	0.23	-0.14	1.18	0.12	11	79
14	1.66	0.29	-0.32	1.69	0.19	10	80
15	1.39	0.24	-1.66	1.41	1.18	10	80
16	0.79	0.14	-0.65	0.80	0.81	10	80
17	1.08	0.19	-1.50	1.10	1.36	10	80
18	1.18	0.15	-0.46	1.19	0.39	7	83
19	0.81	0.16	-0.07	0.83	0.09	11	79
20	0.70	0.11	-0.74	0.71	1.04	9	81
21	0.21	1.32	0.66	1.34	-0.49	81	9
22	0.27	1.19	-0.94	1.22	0.77	77	13
23	0.29	1.62	0.20	1.65	-0.12	80	10
24	0.15	1.20	0.86	1.21	-0.71	83	7
25	0.19	0.98	-1.09	1.00	1.09	79	11
26	0.15	0.77	0.12	0.78	-0.16	79	11
27	0.21	1.53	1.97	1.54	-1.28	82	8
28	0.15	0.97	0.35	0.98	-0.36	81	9
29	0.26	1.45	-0.90	1.47	0.61	80	10
30	0.13	1.09	-0.23	1.10	0.21	83	7
31	0.17	0.86	0.18	0.88	-0.20	79	11
32	0.28	1.59	-0.37	1.61	0.23	80	10
33	0.18	0.84	0.18	0.86	-0.21	78	12
34	0.23	1.48	-0.60	1.50	0.40	81	9
35	0.45	2.58	-0.10	2.62	0.04	80	10
36	0.22	1.40	-1.75	1.42	1.23	81	9
37	0.32	1.38	-0.62	1.42	0.44	77	13
38	0.17	1.05	-0.25	1.06	0.24	81	9
39	0.14	0.89	0.63	0.90	-0.70	81	9
40	0.28	1.42	0.44	1.45	-0.30	79	11

Table 2-4. Item parameters for 20 items with complex structure

Item	a_1	a_2	d	MDISC	MDIFF	α_1	α_2
1	1.12	0.18	-0.70	1.13	0.62	9	81
2	2.23	0.52	-0.23	2.29	0.10	13	77
3	1.39	0.24	-2.19	1.41	1.55	10	80
4	1.02	0.16	-0.75	1.03	0.73	9	81
5	1.68	0.33	-1.18	1.71	0.69	11	79
6	0.22	0.96	0.77	0.99	-0.78	77	13
7	0.20	1.24	1.12	1.26	-0.89	81	9
8	0.16	0.94	-1.26	0.95	1.33	80	10
9	0.20	1.67	-1.86	1.68	1.11	83	7
10	0.29	2.04	-1.46	2.06	0.71	82	8
11	0.99	0.89	-0.72	1.33	0.54	42	48
12	0.55	0.95	0.17	1.10	-0.15	60	30
13	1.26	1.40	0.51	1.89	-0.27	48	42
14	0.49	0.40	-0.07	0.63	0.11	39	51
15	0.60	0.80	-1.38	1.00	1.38	53	37
16	0.87	0.68	-2.66	1.10	2.42	38	52
17	0.83	0.83	-0.48	1.17	0.41	45	45
18	0.68	0.51	-0.68	0.85	0.80	37	53
19	1.48	1.89	-0.77	2.40	0.32	52	38
20	0.56	0.90	-0.73	1.06	0.69	58	32

Table 2-5. Item parameters for 40 items with complex structure

Item	a_1	a_2	d	MDISC	MDIFF	α_1	α_2
1	2.29	0.44	1.28	2.33	-0.55	11	79
2	1.10	0.21	0.21	1.12	-0.19	11	79
3	1.44	0.20	-1.45	1.45	1.00	8	82
4	0.57	0.09	0.28	0.58	-0.48	9	81
5	0.92	0.16	-1.59	0.93	1.71	10	80
6	0.96	0.20	0.45	0.98	-0.46	12	78
7	1.18	0.23	-0.79	1.20	0.66	11	79
8	1.12	0.16	-1.07	1.13	0.95	8	82
9	0.98	0.19	0.80	1.00	-0.80	11	79
10	1.90	0.30	-0.94	1.92	0.49	9	81
11	0.09	0.56	0.27	0.57	-0.48	81	9
12	0.31	1.34	0.40	1.38	-0.29	77	13
13	0.20	1.16	-0.14	1.18	0.12	80	10
14	0.21	1.68	-0.32	1.69	0.19	83	7
15	0.27	1.38	-1.66	1.41	1.18	79	11
16	0.15	0.79	-0.65	0.80	0.81	79	11
17	0.15	1.09	-1.50	1.10	1.36	82	8
18	0.19	1.18	-0.46	1.19	0.39	81	9
19	0.14	0.82	-0.07	0.83	0.09	80	10
20	0.09	0.70	-0.74	0.71	1.04	83	7
21	1.00	0.90	0.66	1.34	-0.49	42	48
22	0.61	1.06	-0.94	1.22	0.77	60	30
23	1.10	1.23	0.20	1.65	-0.12	48	42
24	0.94	0.76	0.86	1.21	-0.71	39	51
25	0.60	0.80	-1.09	1.00	1.09	53	37
26	0.61	0.48	0.12	0.78	-0.16	38	52
27	1.09	1.09	1.97	1.54	-1.28	45	45
28	0.78	0.59	0.35	0.98	-0.36	37	53
29	0.91	1.16	-0.90	1.47	0.61	52	38
30	0.58	0.93	-0.23	1.10	0.21	58	32
31	0.61	0.63	0.18	0.88	-0.20	46	44
32	1.22	1.06	-0.37	1.61	0.23	41	49
33	0.49	0.70	0.18	0.86	-0.21	55	35
34	1.35	0.66	-0.60	1.50	0.40	26	64
35	2.04	1.65	-0.10	2.62	0.04	39	51
36	1.07	0.93	-1.75	1.42	1.23	41	49
37	1.04	0.97	-0.62	1.42	0.44	43	47
38	0.88	0.59	-0.25	1.06	0.24	34	56
39	0.42	0.79	0.63	0.90	-0.70	62	28
40	1.11	0.93	0.44	1.45	-0.30	40	50

CHAPTER 3
RESULTS

As described in Chapter 2, the criterion used in this study to evaluate the four multidimensional IRT linking methods was based on the differences between the item parameter estimates for the base group and the transformed item parameter estimates for the equated group across 500 replications. Specifically, after the item parameter estimates from the two groups were transformed to a common scale, the mean and standard deviation of their differences across the 500 replications were computed to examine the accuracy and stability of the four linking methods. For each of the 192 experimental conditions, there were three parameter estimates a_1, a_2, and d; therefore, the mean and standard deviation of the differences were computed for a_1, a_2, and d across 500 replications for each item of the test. Then the distributions of the means and standard deviations of the differences of a_1, a_2, and d for all items in the test were obtained. Based on the characteristic of item parameter invariance in IRT, the item parameter estimates from the base and equated groups should theoretically be equal after they are transformed to a common scale. So their differences, and accordingly the means and standard deviations of their differences across 500 replications, should be 0. Therefore, the performance of the four multidimensional IRT linking methods can be evaluated by examining how close the means and standard deviations of the differences are to 0.

There is currently no generally accepted criterion about how close the item parameter estimates for the two groups should be in order for the linking to be considered accurate and stable. To describe the distribution of difference, histograms of means and standard deviations of the differences of a_1, a_2, and d for the 192 experimental conditions were prepared. The appendix contains the histograms for all 192 conditions. In this chapter, histograms selected to

illustrate the trends in the results will be presented. The following midpoints were used to construct the histograms for means: 0, ±0.2, ±0.4, ±0.6. All values smaller than -0.5 and larger than +0.5 were included in the categories with midpoints of ±0.6. For the histograms of the standard deviations, 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, and 1.3 were used as the midpoints. All values beyond 1.2 were classified into the category with midpoints of 1.3. If all or most of the items in the test had means and standard deviations close to 0, the linking method was considered accurate and stable. Otherwise, the linking method was inaccurate and unstable. The performance of the four multidimensional IRT linking methods was evaluated in this way under different testing conditions.

On the histograms, the direct method, equated function method, test characteristic function method, and item characteristic function method are labeled Link1, Link2, Link3, and Link4. For test structure, the approximate simple structure is abbreviated as APP and the complex structure as COM. For test length, the number of items in the test is indicated by n = 20 or n = 40. For sample size, the number of examinees is indicated by N = 500, N = 1000, or N = 2000. For ability distribution differences between the base and equated groups, the condition is abbreviated as G1 if the mean vectors and covariance matrices were equal for the two groups, G2 if only the mean vectors were different, G3 if only the covariance matrices were different, and G4 if the mean vectors and covariance matrices were not equal for the two groups.

As will be shown subsequently, inspection of the results indicated that the effects of linking methods depended on the test structure. Therefore, the decision was made to focus primarily on the effects of linking methods within each of the test structures. Inspection of the results for APP suggested that the interactions of all other factors were small in size, therefore the focus was on the main effects of the factors. Inspection of the COM results suggested that

there were two-way, three-way, or four-way interactions of other factors, so the performance of the linking methods were described taking into account these interactions.

This chapter consists of six sections. The first section compares the general performance of the four linking methods. The second section compares the four linking methods for different test structures. The third section compares linking methods for tests with different lengths. The fourth section compares linking methods for different sample sizes. The fifth section compares linking methods for groups with different ability distributions. The last section shows the relationship between scale linking performance and item parameter values.

General Performance of the Different Linking Methods

The performance of the four multidimensional IRT linking methods was first compared across all testing conditions by collapsing the means and standard deviations of a_1, a_2, and d for all items under different testing conditions. The histograms in Figure 3-1 show the distributions of means and standard deviations for a_1, a_2, and d across all items and both test structures. Based on the percentage of items with the means and standard deviations of differences close to 0, Link1 (direct method) produced more accurate and stable linking results than Link4 (item characteristic function method), and Link4 yielded more accurate and stable results than Link3 (test characteristic function method). Link2 (equated function method) did not provide accurate and stable results for a high percentage of items.

The performance of the four linking methods was also examined separately for different test structures. Figure 3-2 shows the distributions of means and standard deviations for a_1, a_2, and d for APP and COM conditions. Comparing the histograms for the four linking methods on the left side of the figures, one can see that there was no apparent difference among the four linking methods for APP conditions. The histograms on the right side of the figures show that

there was obvious difference among the four linking methods for COM conditions. Specifically, based on the accuracy and stability of linking function, (a) Link1 (direct method) worked well, (b) Link2 (equated function method) worked poorly, and (c) the performance of Link3 (test characteristic function method) and Link4 (item characteristic function method) was between that of Link1 and Link2, with Link4 being slightly better than Link3.

In sum, Link1 (direct method) was consistently the best method and Link2 (equated function method) the worst method under most COM conditions; the four linking methods worked equally and consistently well under most APP conditions.

Performance of Linking Methods for Different Test Structures

In this section, the performance of the four linking methods is compared between APP and COM conditions. Figure 3-2 shows different linking results for the two test structures. Based on the histograms for APP and COM conditions in the figure, all the four linking methods produced more accurate and more stable results for APP tests than for COM tests, but the difference in quality of linking varied across the linking methods. Specifically, Link1 (direct method) results were slightly better for APP tests than for COM tests, especially for parameters a_1 and a_2; Link3 (test characteristic function method) and Link4 (item characteristic function method) results were much better for APP tests than for COM tests; Link2 (equated function method) yielded very poor results for COM tests, but good results for APP tests.

However, for the large sample size (N = 2000), Link1 (direct method), Link3 (test characteristic function method), and Link4 (item characteristic function method) worked almost equally well for APP tests and COM tests; the linking performance difference between APP and COM conditions still remained for Link2 (equated function method) due to its poor function for COM conditions. Figure 3-3 shows the results of four linking methods for APP and COM tests

when the sample size is 2000. With smaller sample sizes (N = 500 and N = 1000), the linking performance difference between APP and COM conditions increased for Link3, Link4, and Link1 (see specific histograms in the appendix).

Therefore, test structure had its smallest effects on Link1 (direct method), larger effects on Link3 (test characteristic function method) and Link4 (item characteristic function method), and the largest effect on Link2 (equated function method). Link2 worked well for all APP tests, but poorly for all COM tests in this study. Due to the strong influence of test structure on the function of the four linking methods, most of the results in the following sections are presented separately for the APP and COM conditions.

Performance of Linking Methods for Different Test Lengths

Given the different performance of linking methods for APP and COM conditions, the influence of test lengths on the linking function was explored separately for APP and COM tests. The distributions of means and standard deviations of differences for a_1, a_2, and d for short and long tests under the APP conditions are presented in Figure 3-4. Based on the histograms in the figure, one can see that for APP tests, although the linking performance was not strongly influenced by test length, all four linking methods produced slightly more accurate and stable results with long tests.

Inspection of the results indicated that under COM conditions, the performance of the linking methods depended on the sample size and test length. Therefore, the influence of test lengths was next explored separately for different sample sizes for COM tests. Figure 3-5 shows the linking results for short and long tests with sample size of 500. Although none of the four linking methods worked well, the histograms still show that the results for Link1 (direct method), Link3 (test characteristic function method), and Link4 (item characteristic function method) for

short tests were better than those for long tests and that Link1 to some extent performed similarly for different test lengths.

Figure 3-6 illustrates the linking results for short and long tests with sample size of 1000. From the figure, it was difficult to state at which test length linking performance was better. Subsequently reported results will show that the performance depended to some extent on the ability distribution difference between the base and equated groups. When the linking results for ability condition 2 (G2: unequal mean vectors) were excluded, the linking function for long test was obviously better than that for short test except for Link2 (equated function method), as presented in Figure 3-7. Therefore, in Figure 3-6, the performance under G2 masks the positive effect of test length on the linking accuracy and stability.

Shown in Figure 3-8 are linking results for different test lengths with sample size of 2000. It is very obvious that the linking results for long tests were better than those for short tests except for Link2 (equated function method).

In sum, the linking results for long tests were better than those for short tests except in some COM conditions when the sample size was small.

Performance of Linking Methods for Different Sample Sizes

Inspection of results indicated that sample size had stable and consistent influence on the linking performance, but with different degrees of influence for different test structures. Therefore, the effect of sample size is first shown across all other testing conditions then presented separately for APP and COM conditions. Figure 3-9 contains the linking results for different sample sizes. Comparing horizontally the histograms for different sample sizes, one can find that both the linking accuracy and stability increased for Link1 (direct method), Link3 (test characteristic function method), and Link4 (item characteristic function method) with the sample sizes changing from 500 and 1000 to 2000. However, the linking performance increased at

different degrees for different test structures. Figure 3-10 shows the linking results for different sample sizes for APP tests. Figure 3-11shows the results for different sample sizes for COM tests., From Figure 3-10, it can be found that the accuracy of the four linking methods was fairly good at all sample sizes and the stability of the four linking methods increased when the sample size became large for APP tests. Figure 3-11 suggests that although the accuracy and stability of Link1, Link3, and Link4 increased when the sample size became large for COM tests, the linking performance was poor for sample sizes of 500 and 1000 especially for Link3 and Link4. In addition, for COM tests, the accuracy and stability for Link2 (equated function method) were very poor for all sample sizes and relatively unaffected by sample size.

Based on these findings, (a) Link1 (direct method), Link3 (test characteristic function method), and Link4 (item characteristic function method) for APP tests were less affected by different sample sizes than were COM tests; (b) Link1 (direct method) was less affected by sample sizes than were the other linking methods for COM tests.

Performance of Linking Methods for Groups with Different Ability Distributions

Inspection of the results suggested that the linking results for different ability distributions depended on other testing conditions. Therefore, the influence of ability distribution was first explored separately for APP and COM tests. Figure 3-12 shows the linking results for groups with different ability distributions under the APP conditions. Comparing horizontally the histograms across G1 (equal mean vectors, equal covariance matrices), G2 (unequal mean vectors, equal covariance matrices), G3 (equal mean vectors, unequal covariance matrices), and G4 (unequal mean vectors, unequal covariance matrices) indicates that: (a) for a_1 and a_2, the linking results for G1 were slightly better than those for other ability conditions; (b) for d, the results for G2 were somewhat worse than those for other ability conditions; (c) Link2 (equated

function method) was least affected by ability distributions. The results imply that a difference between groups in the mean vectors was more influential than a difference between the groups in the covariance matrices.

Inspection of results for COM tests indicated that the influence of ability distribution was moderated by sample size; Therefore, the effect of ability distribution was explored separately for N=500, N=1000, and N=2000 for COM tests with the concentration on Link1 (direct method), Link3 (test characteristic function method), and Link4 (item characteristic function method). Figure 3-13 shows the linking results for groups with different ability distributions with sample size of 500. One can see from the figure that although none of the linking methods worked well for the small sample size, the linking results for G2 (unequal mean vectors, equal covariance matrices) and G4 (unequal mean vectors, unequal covariance matrices) were worse than for G1 (equal mean vectors, equal covariance matrices) and G3 (equal mean vectors, unequal covariance matrices). Even though Link1 (direct method) was relatively unaffected by between-group difference in ability distributions in groups, it still did not work well in linking d for G2, which indicates the strong influence of the mean difference between groups. The linking results for different groups with sample size of 1000 presented in Figure 3-14 shows that linking methods did not work well under G2, especially for d. However, there was some interaction between group differences and test length. For long test (n = 40), the linking methods did not work well for G2 (see Figure 3-15); for short test (n = 20), the linking methods worked relatively well for G2 (see Figure 3-16). The linking results for different groups with sample size of 2000 shown in Figure 3-17 suggest that the linking methods worked approximately equally well for the groups with different ability distributions.

In sum, the influence of ability distributions on linking results depended on other testing conditions: (a) between-group differences in ability distributions did not have a strong influence on the performance of the four linking methods for APP conditions or for COM conditions with a large sample size; (b) mean difference between groups had negative influence on the linking results especially for conditions with small sample size.

Performance of Linking Methods for Test Items with Different Parameter Values

Two types of scatter-plots were used to examine the relationship between linking performance and item parameter values under each of the 48 testing conditions. The first type of scatter-plot was used to evaluate the effect of item parameter values on the accuracy of different linking methods, with y axis as the mean of the differences and x axis as the true parameter values which were used to generate the item response data. The second type of scatter-plots was used to evaluate the effect of item parameter values on the stability of different linking methods, with y axis as the standard deviation of the differences and x axis as the true parameter values. These two scatter-plots were constructed for each of the three parameter estimates (e.g., a_1, a_2, and d) under each of the 48 testing conditions. However, the results of Link2 (equated function method) were not included in the scatter-plots under the COM conditions due to its consistently poor performance. Given the limitation of space, the main outcomes are illustrated by some representative examples.

The results suggest that: (a) Under most of the testing conditions, the linking results tended to be less accurate for a_1 and a_2 when the two parameters had extreme values, and (b) under most of the testing conditions, the linking results became less stable for a_1 and a_2 as the parameters values increased. The results also indicate that: (a) The accuracy of linking results for d was not closely related to their true parameter values under most of the testing conditions, and

(b) the stability of linking results for d was also not closely related to their true parameter values when the sample size was not large. The scatter-plots for one testing condition (COM, n = 20, N = 1000, G3) illustrate these relationships between linking performance and item parameter values (see Figure 3-18).

However, for large sample size (N = 2000) the stability of linking results for d was closely related to their absolute true parameter values. Specifically when the absolute parameter values of d were closer to 0, the linking results were more stable; when the absolute parameter values of d were farther away from 0, the linking results were less stable. The scatter-plots for another testing condition (APP, n = 40, N = 2000, G4) show that: (a) the linking results tended to be less accurate and less stable for a_1 and a_2 when the two parameters had extreme values; (b) the linking results tended to be less stable for d as the absolute parameters values increased (see Figure 3-19).

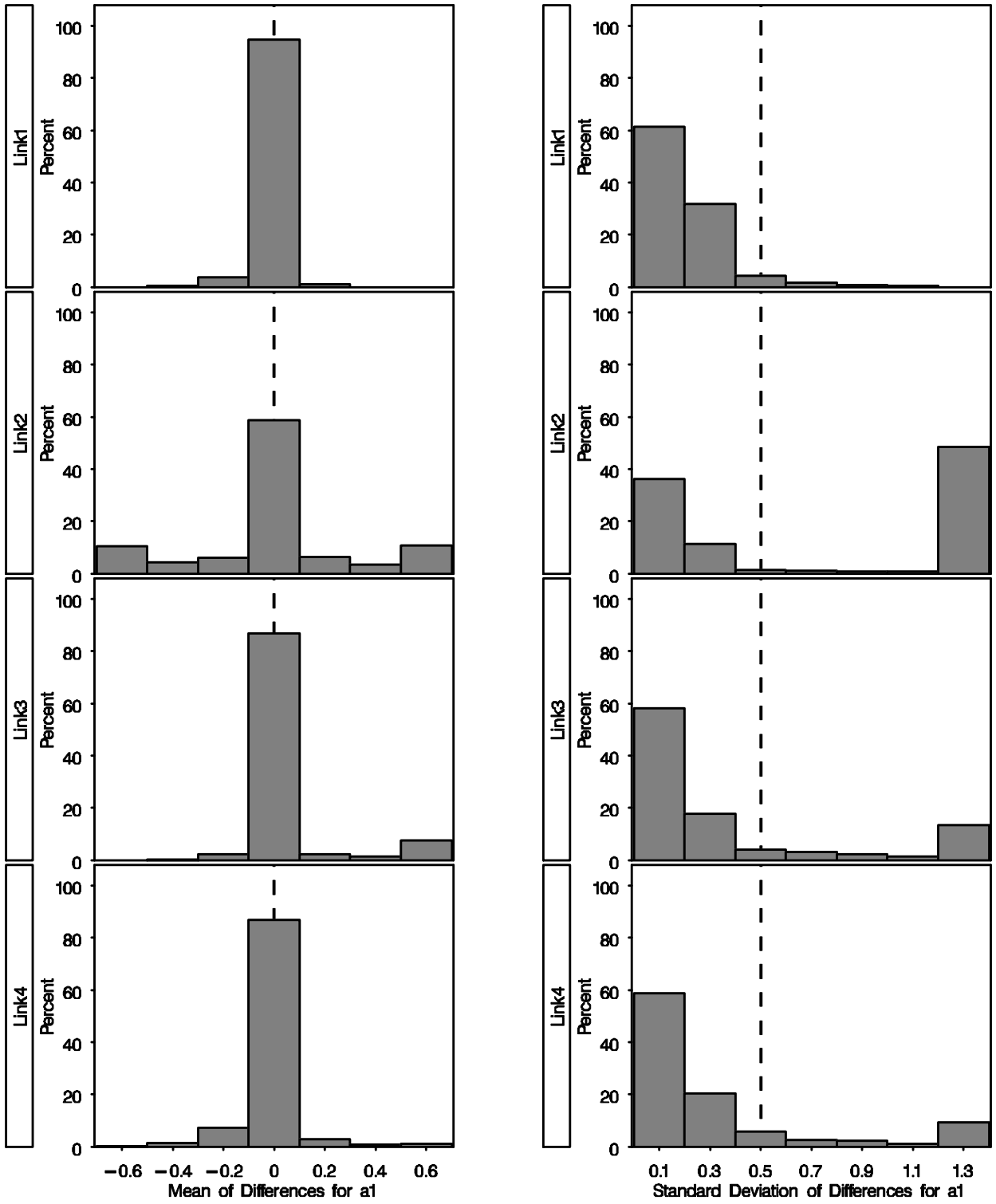


Figure 3-1. Accuracy and stability for different linking methods

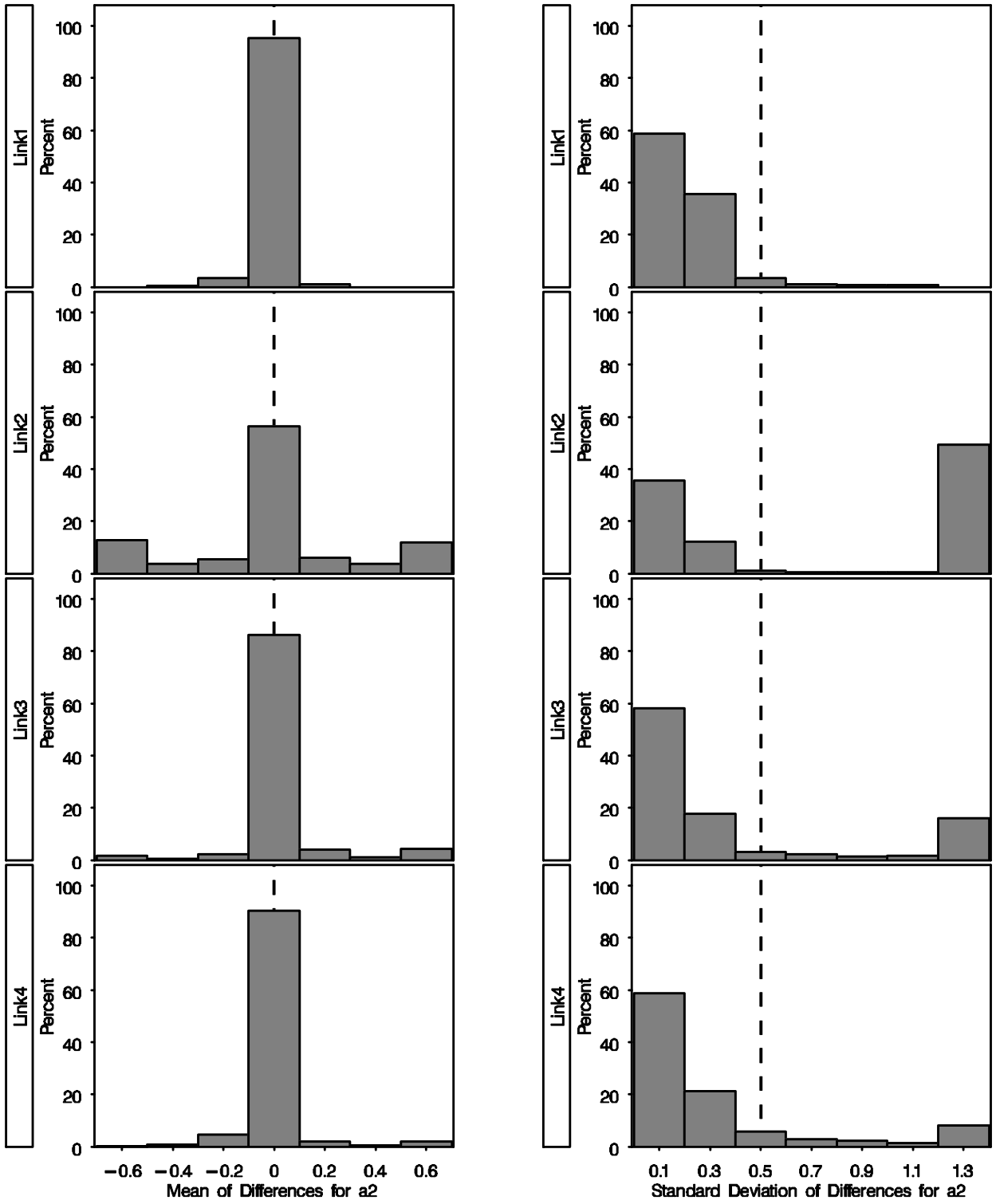


Figure 3-1. Continued

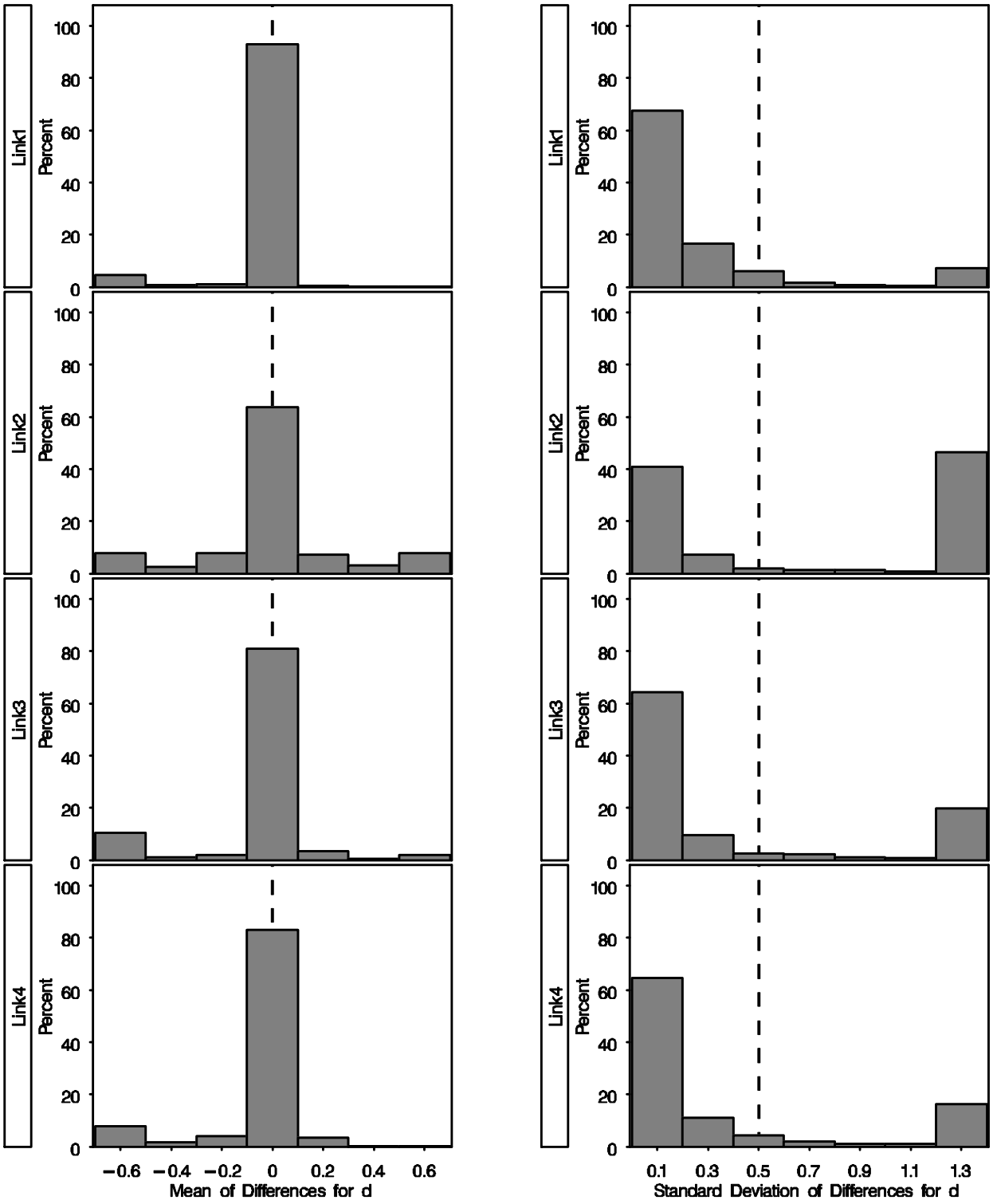


Figure 3-1. Continued

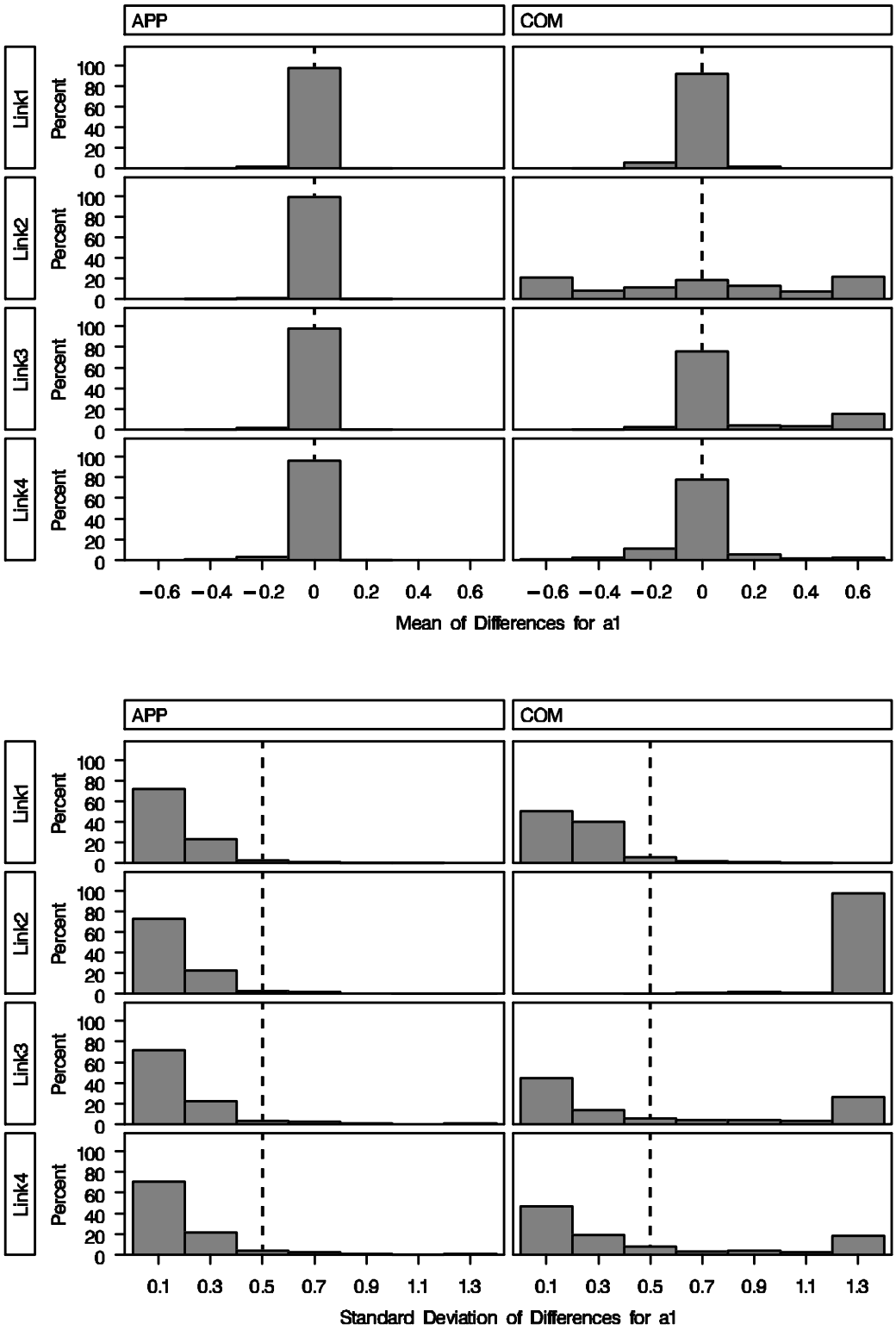
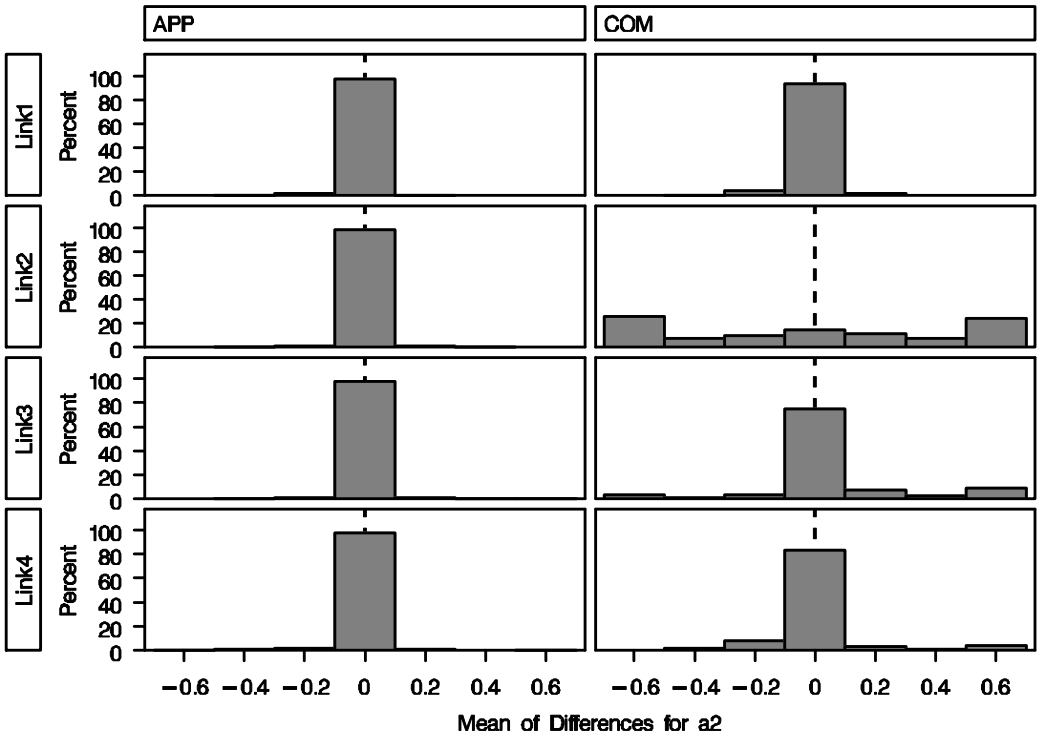


Figure 3-2. Accuracy and stability by linking method and test structure



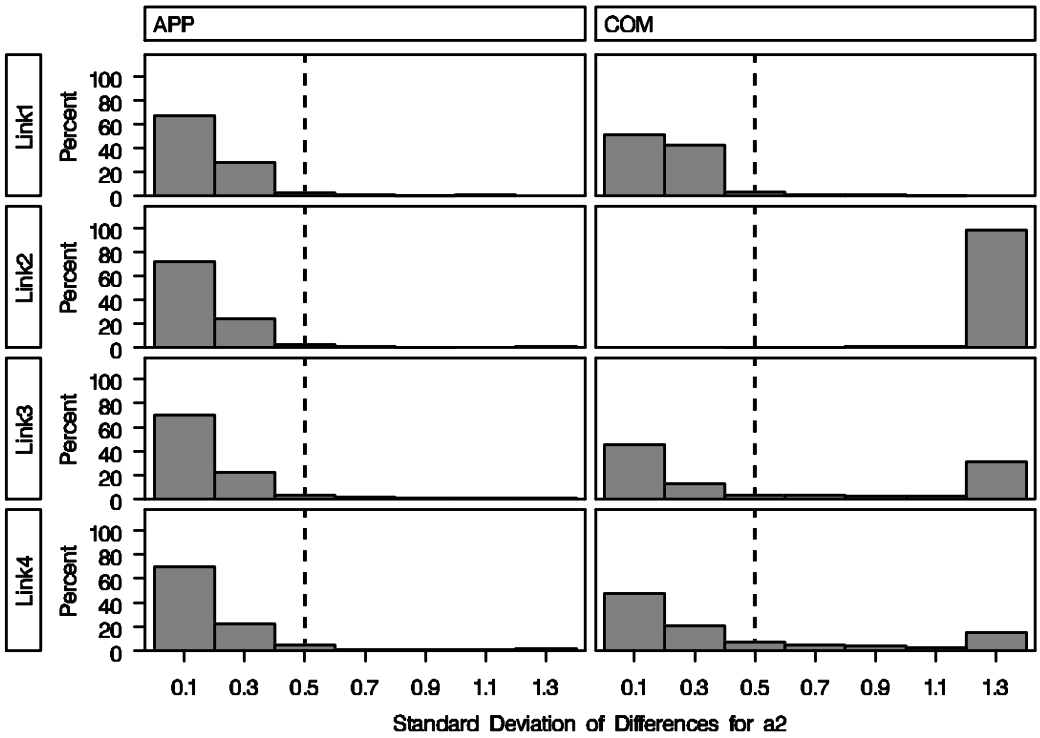


Figure 3-2. Continued

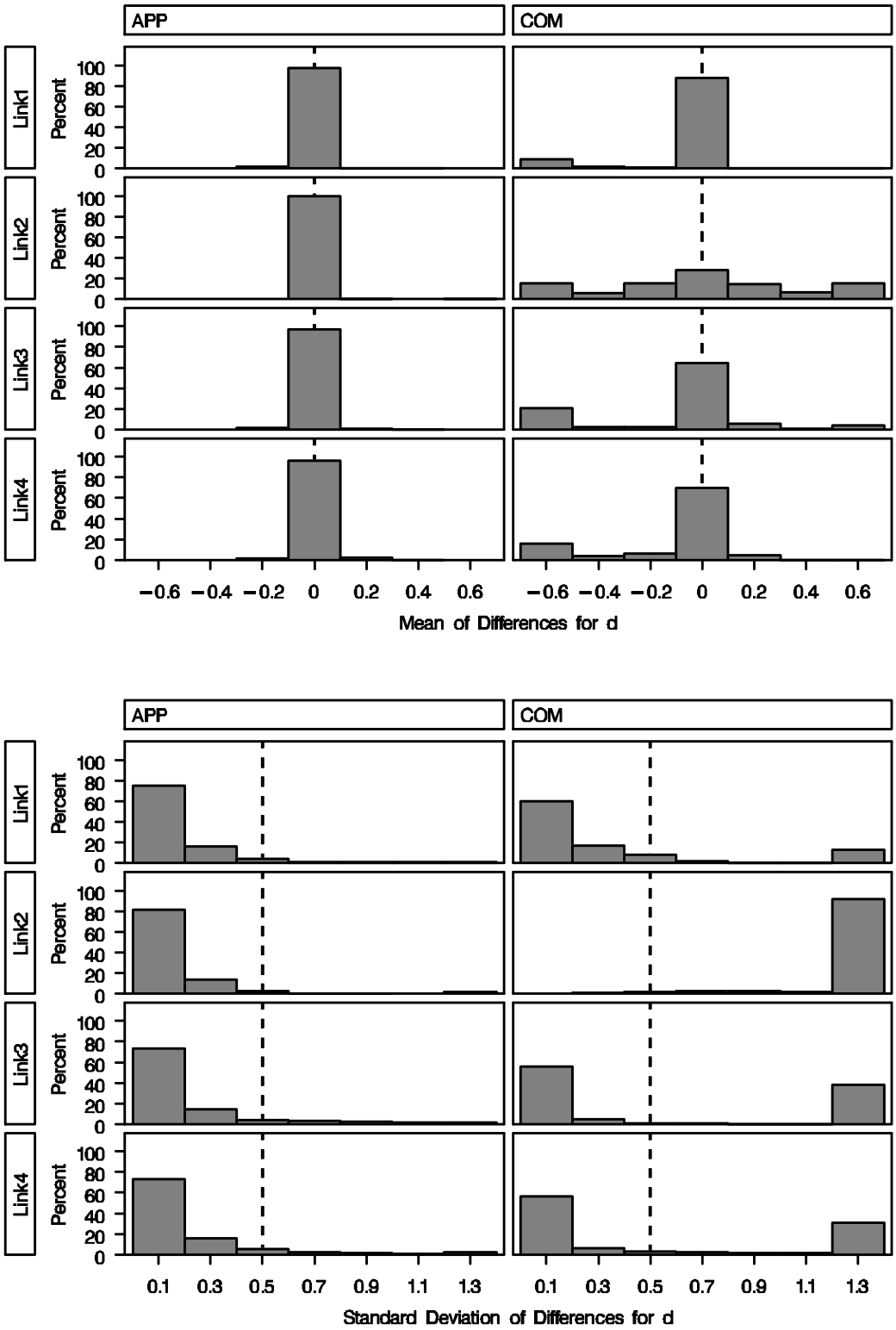


Figure 3-2. Continued

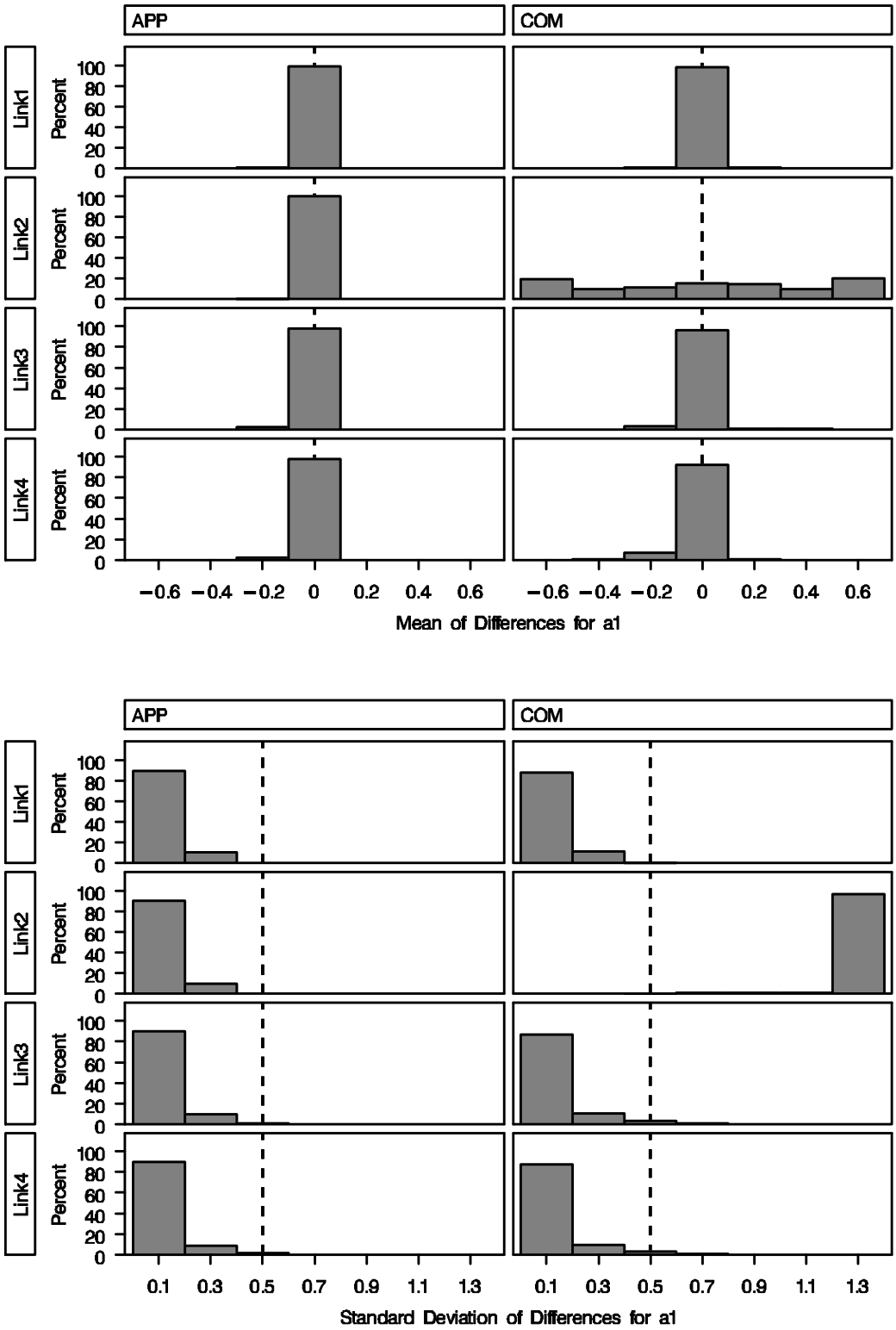


Figure 3-3. Accuracy and stability by linking method and test structure: N = 2000

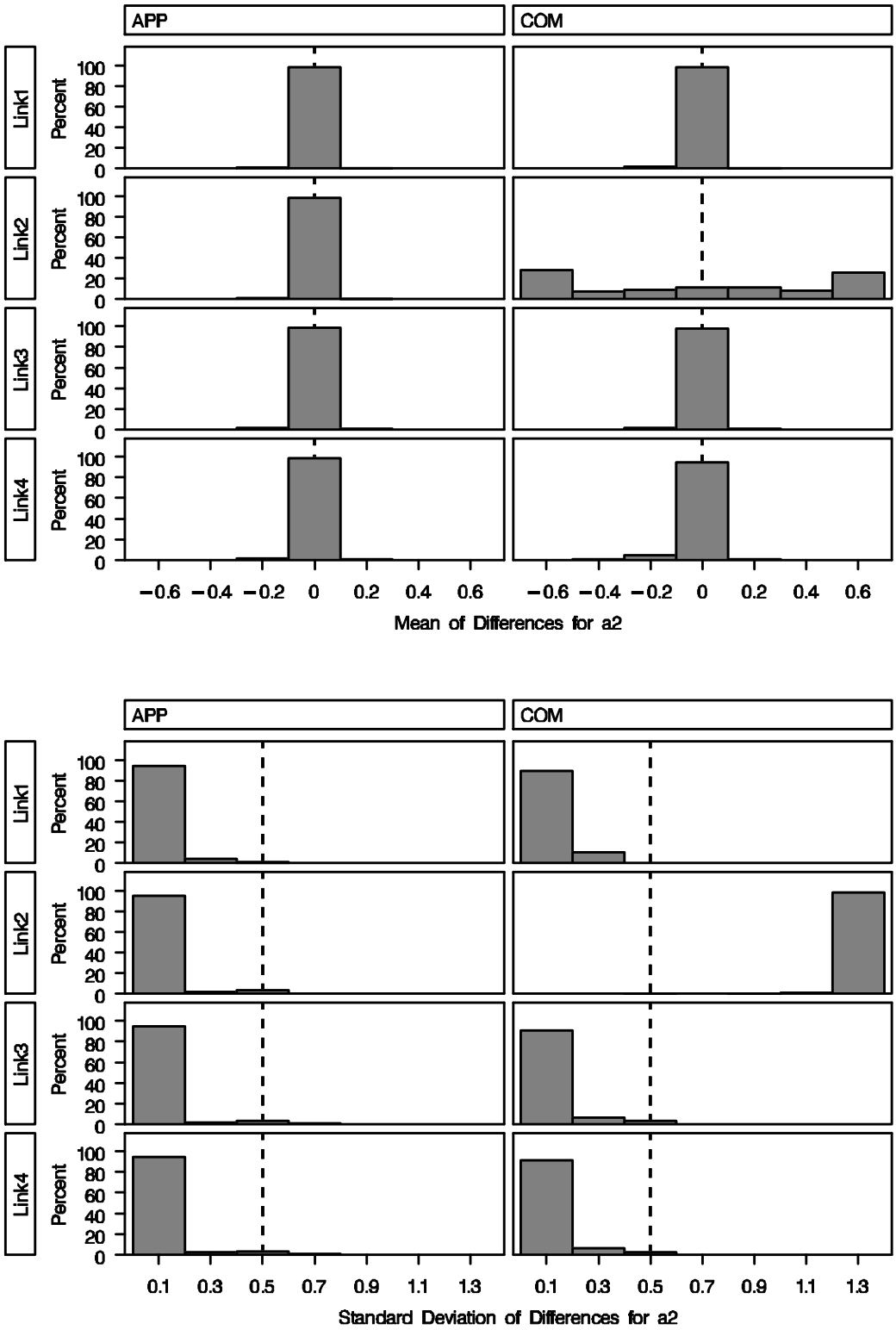


Figure 3-3. Continued

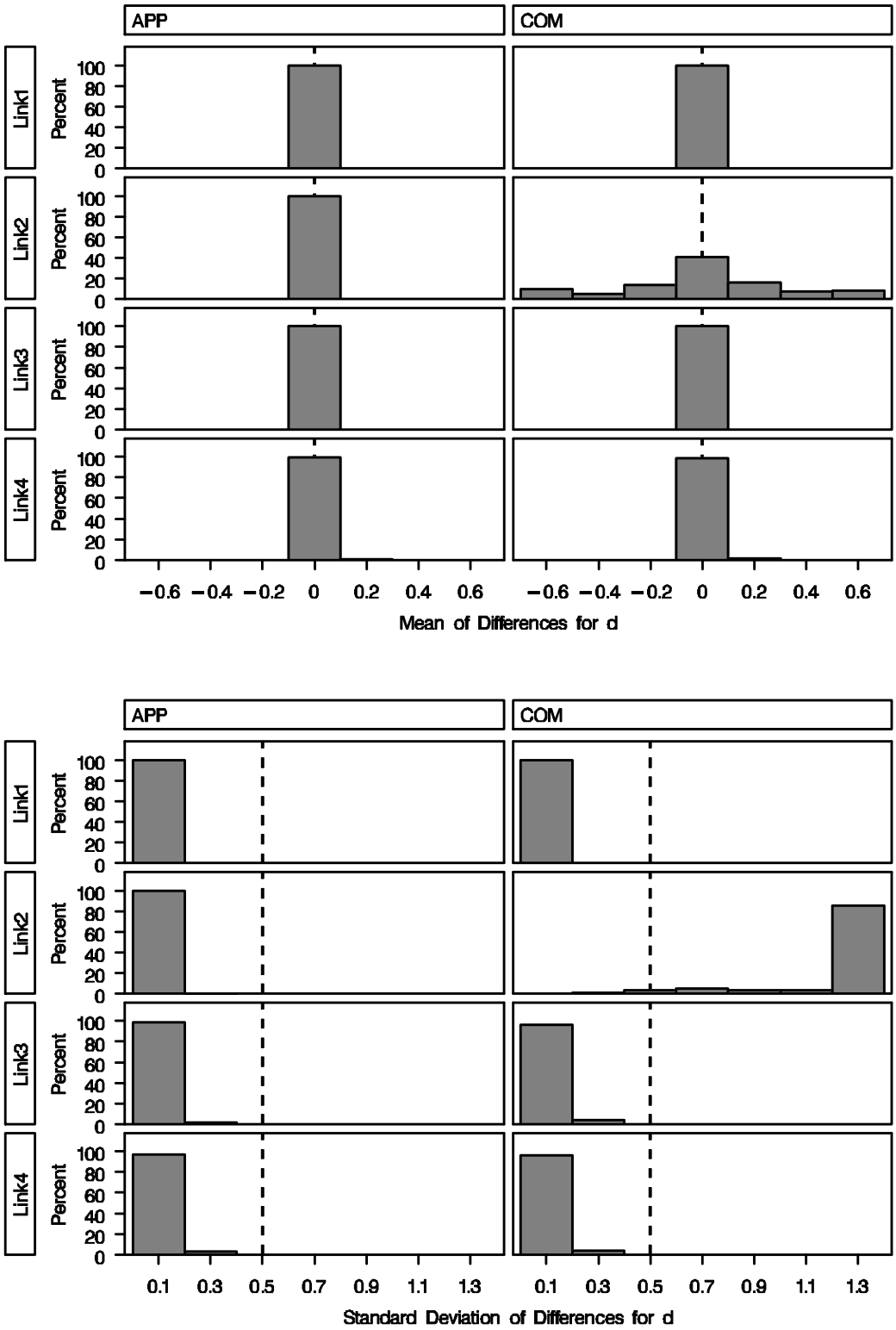


Figure 3-3. Continued

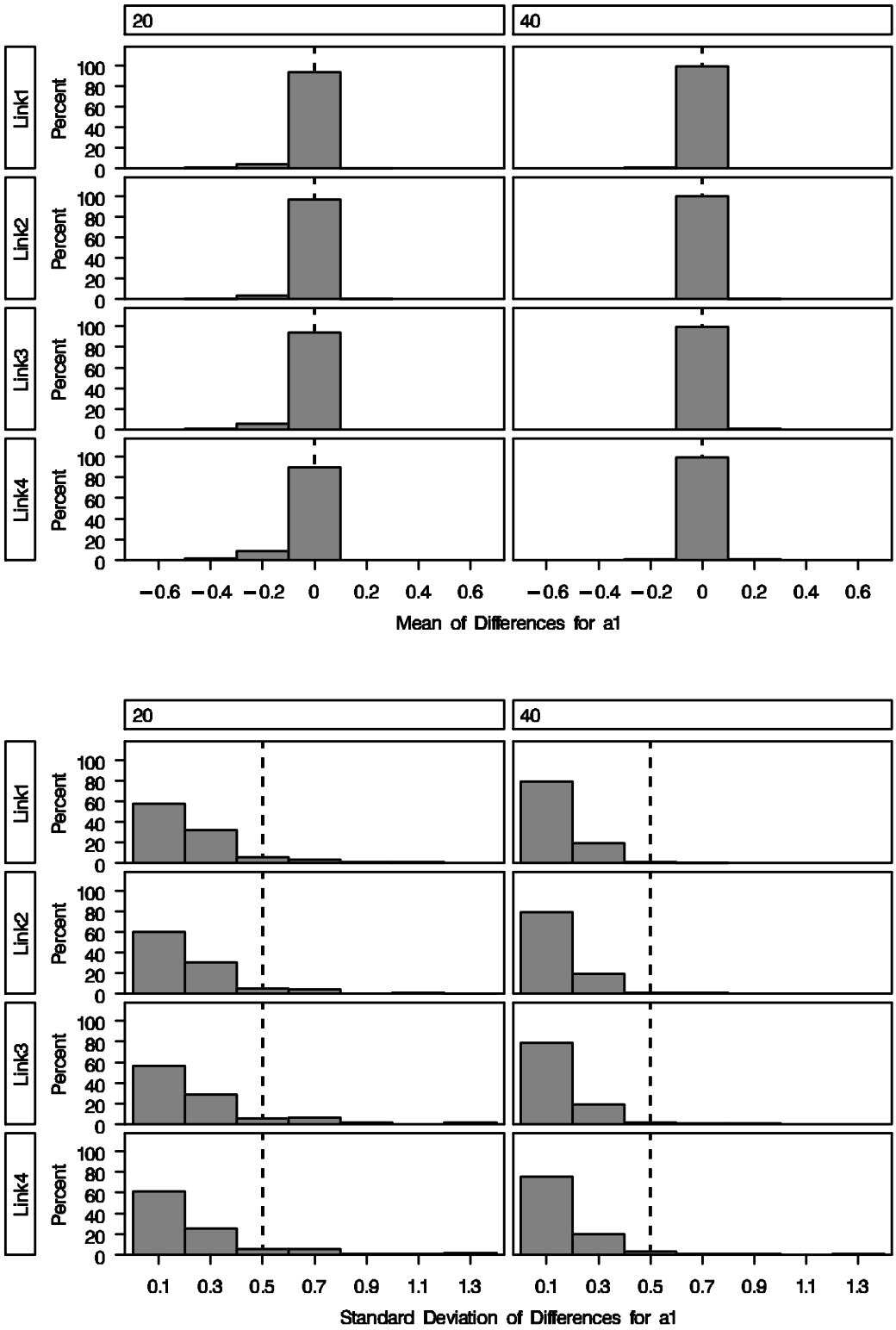


Figure 3-4. Accuracy and stability by linking method and test length for approximate simple structure tests

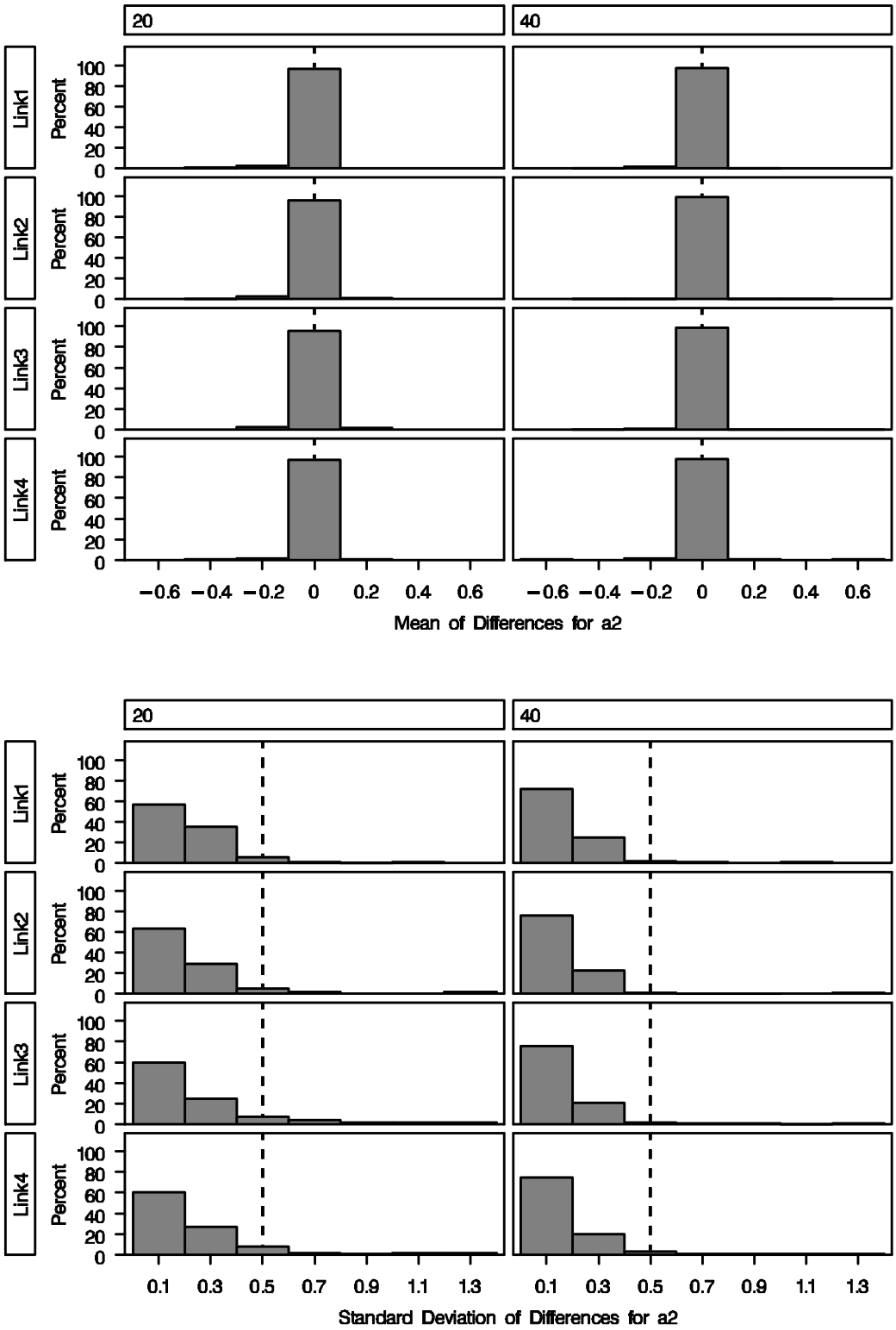


Figure 3-4. Continued

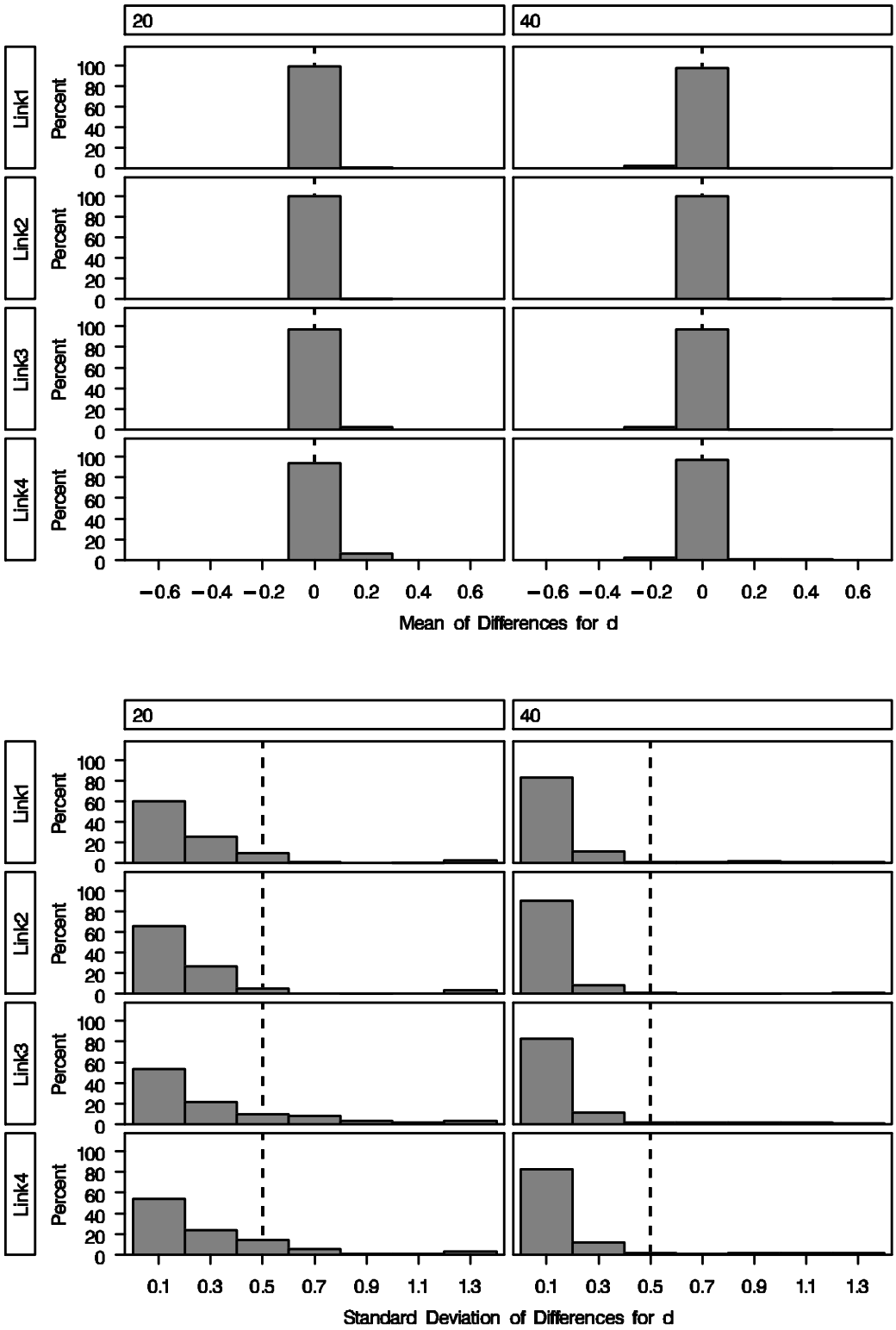


Figure 3-4. Continued

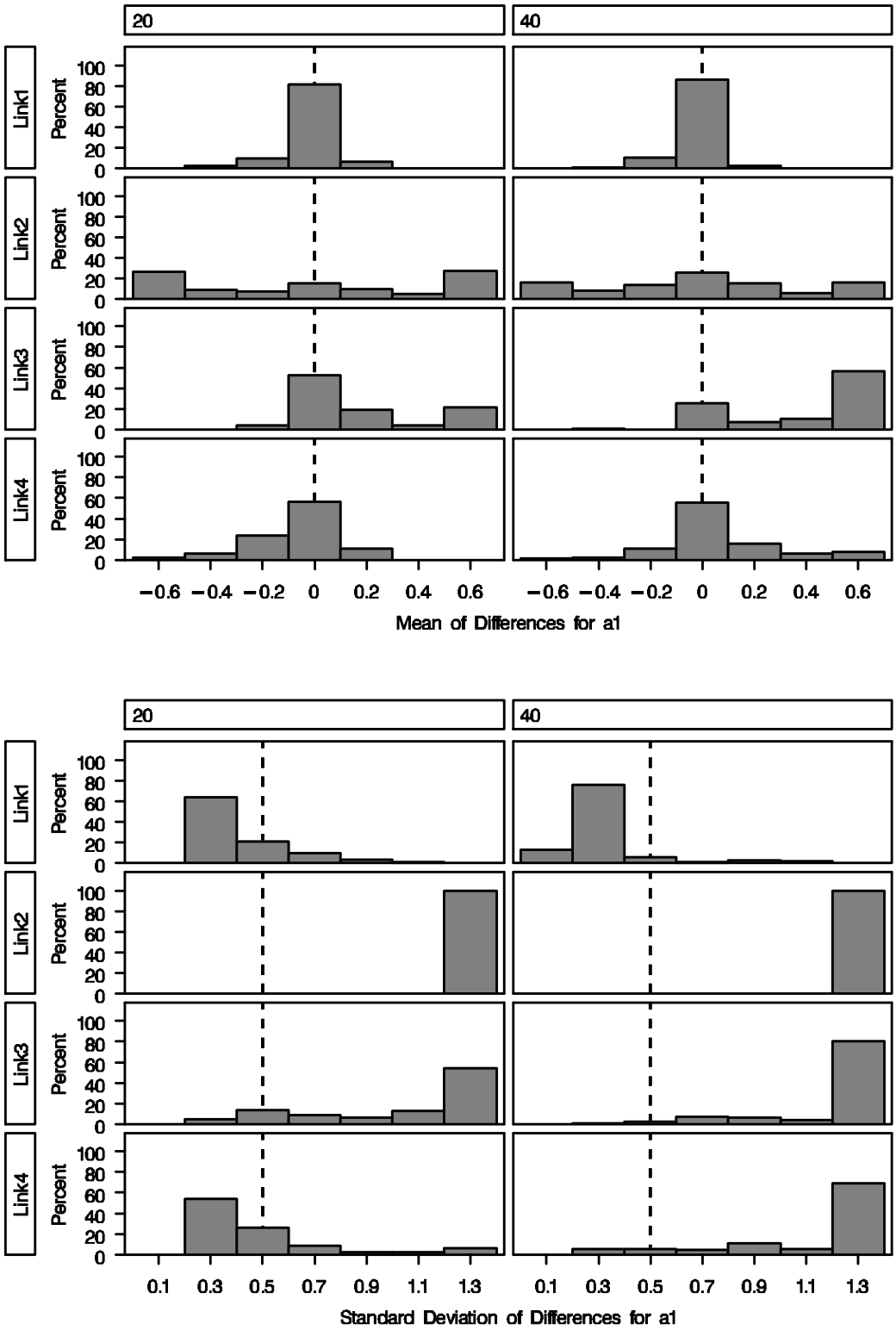


Figure 3-5. Accuracy and stability by linking method and test length for complex structure tests: N = 500

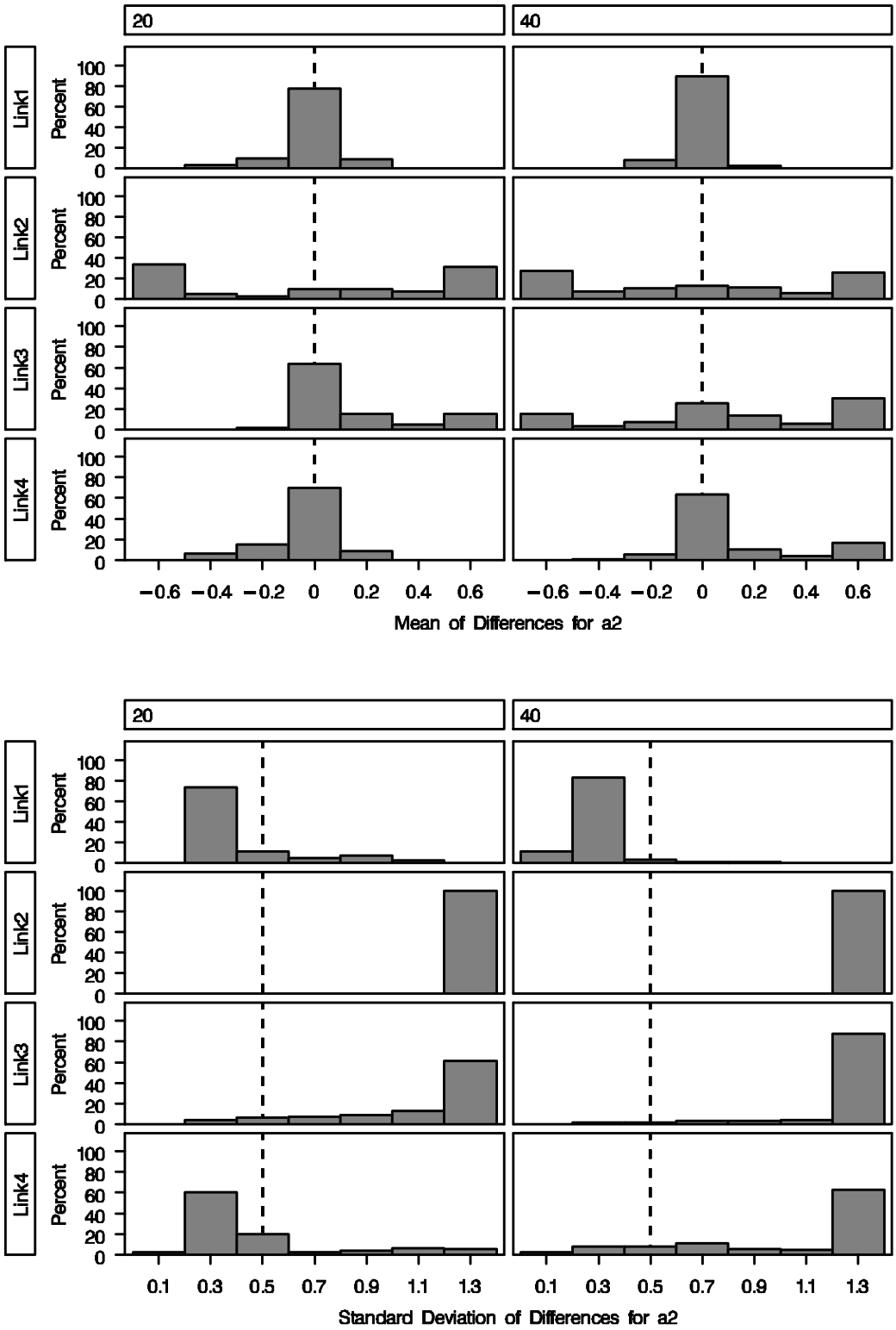


Figure 3-5. Continued

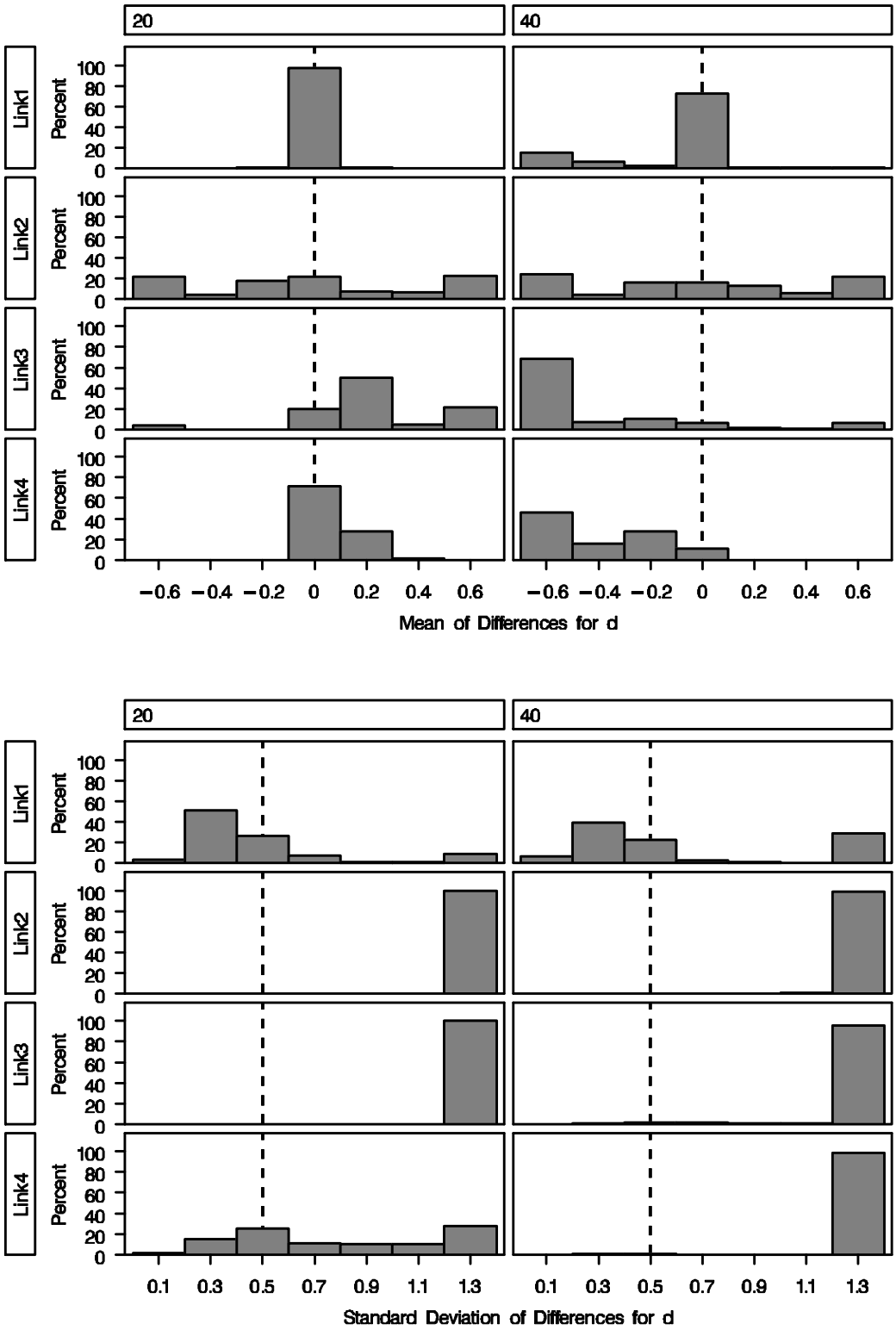


Figure 3-5. Continued

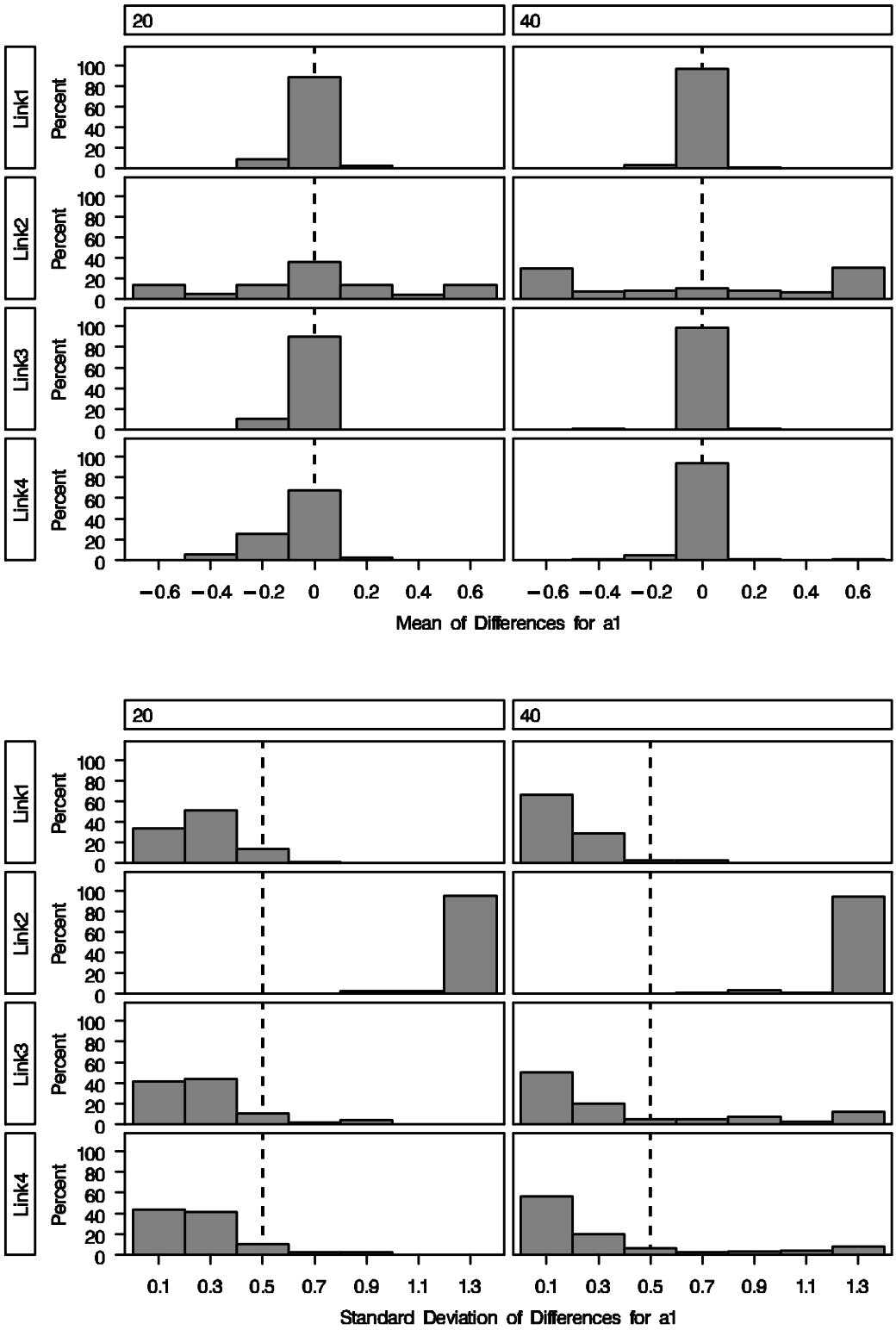


Figure 3-6. Accuracy and stability by linking method and test length for complex structure tests: N = 1000

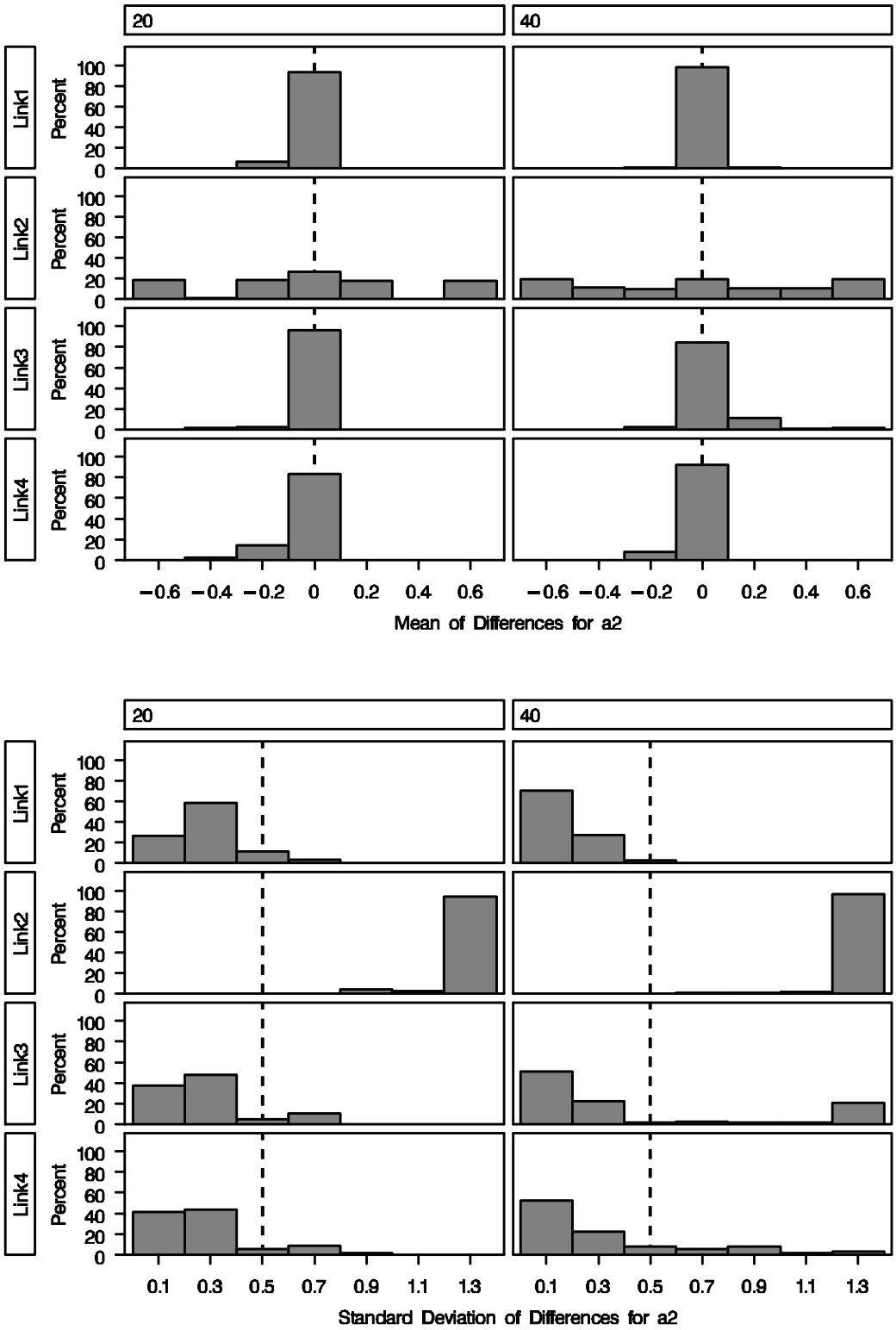


Figure 3-6. Continued

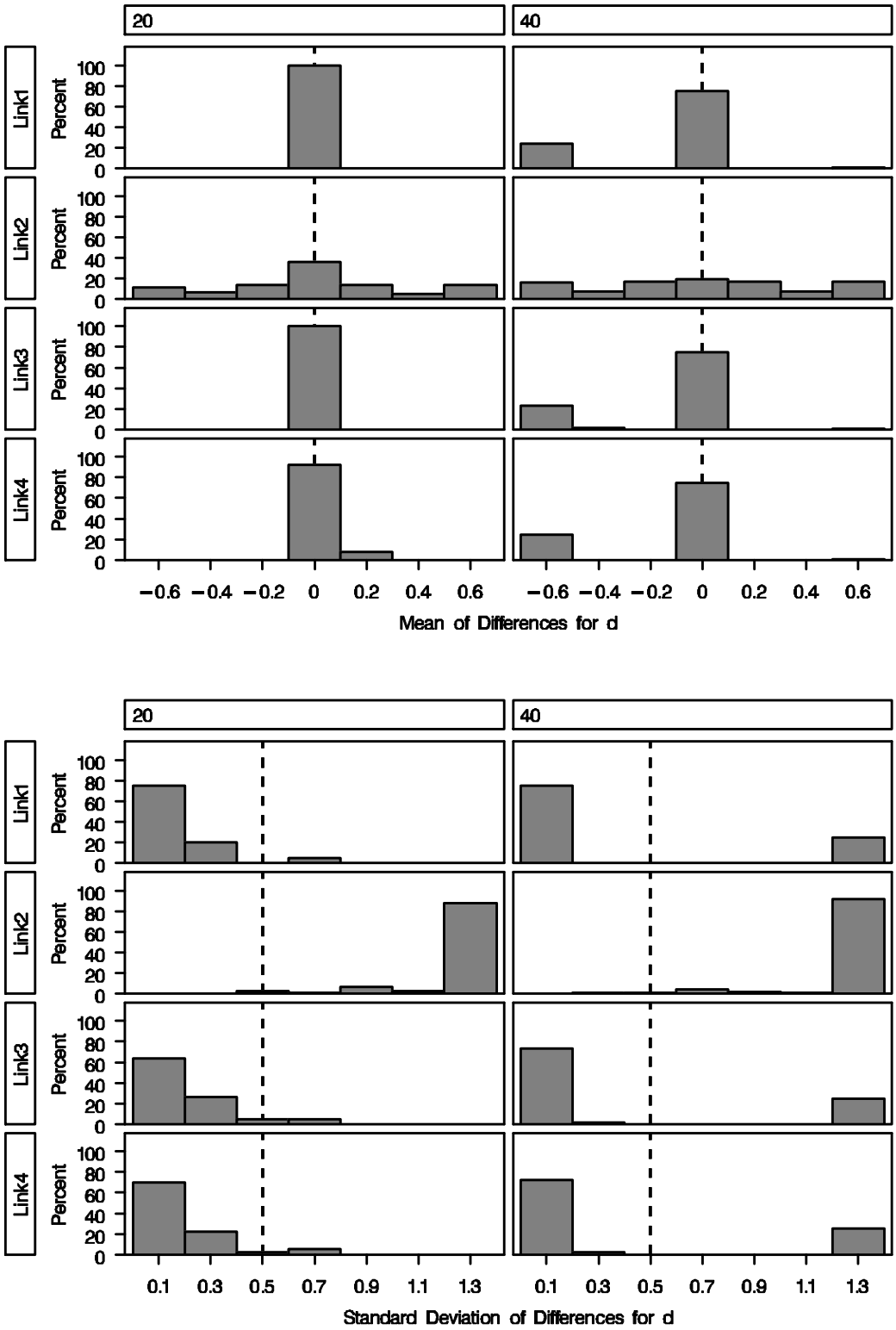


Figure 3-6. Continued

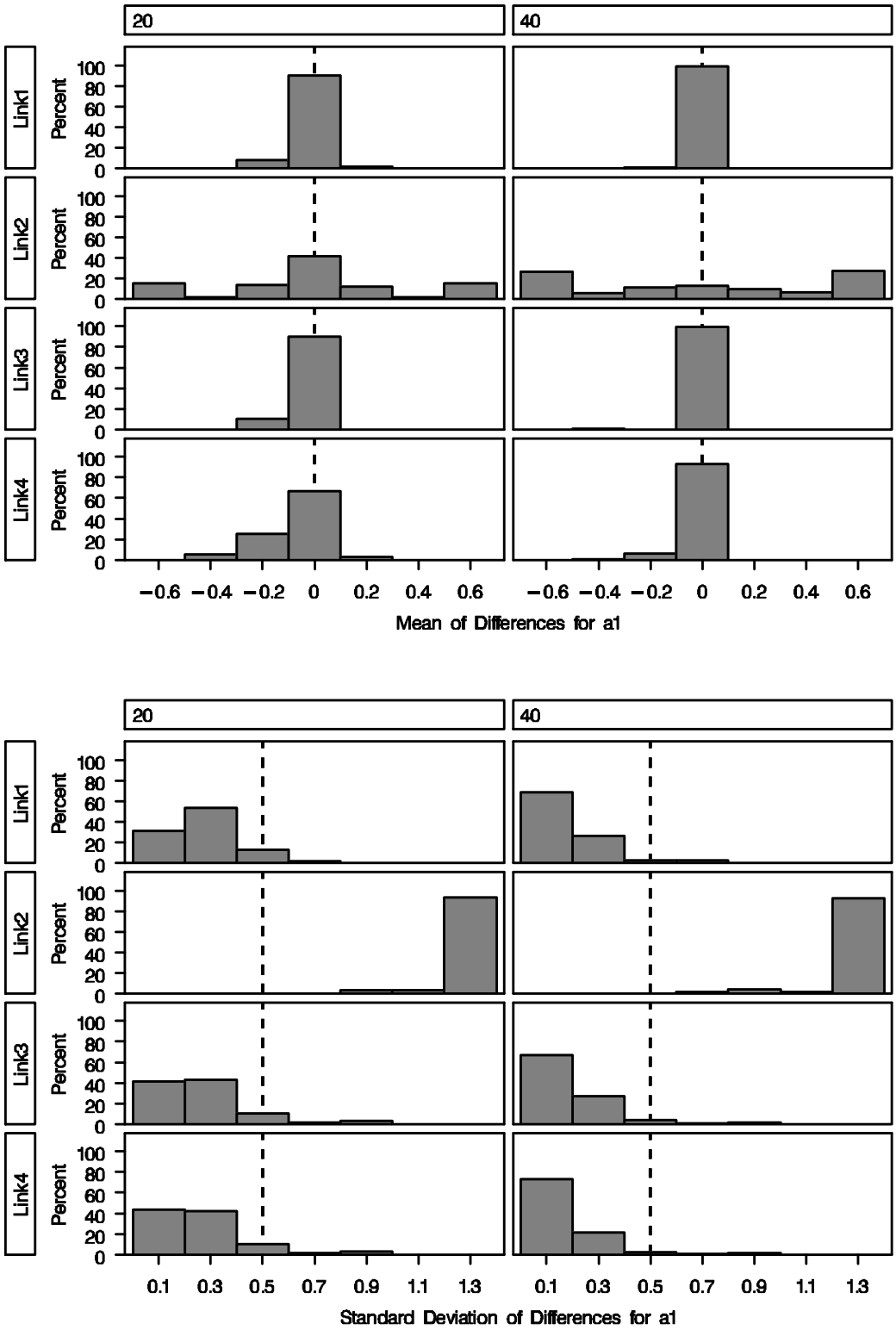


Figure 3-7. Accuracy and stability by linking method and test length for complex structure tests when G2 was excluded: N=1000

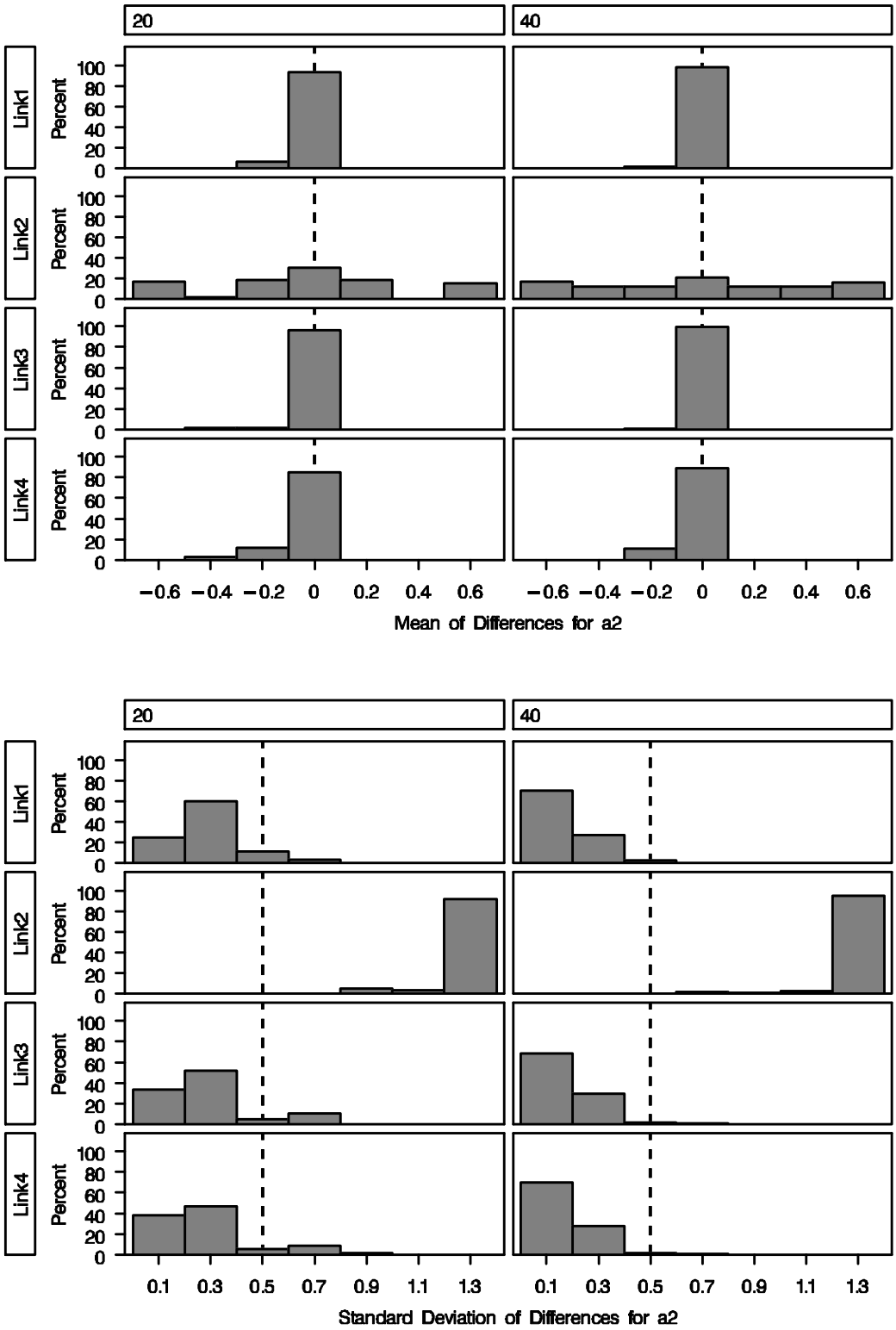


Figure 3-7. Continued

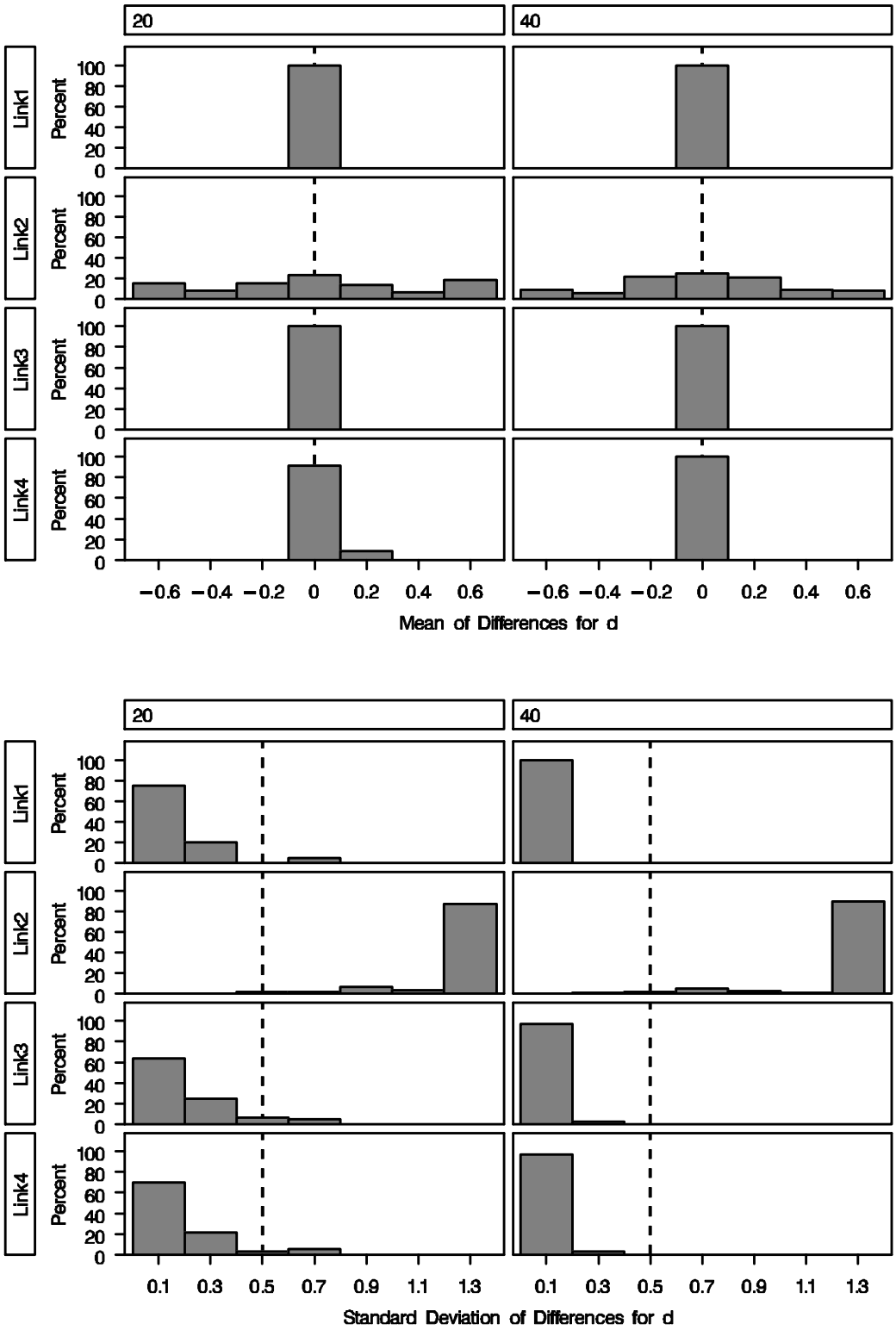


Figure 3-7. Continued

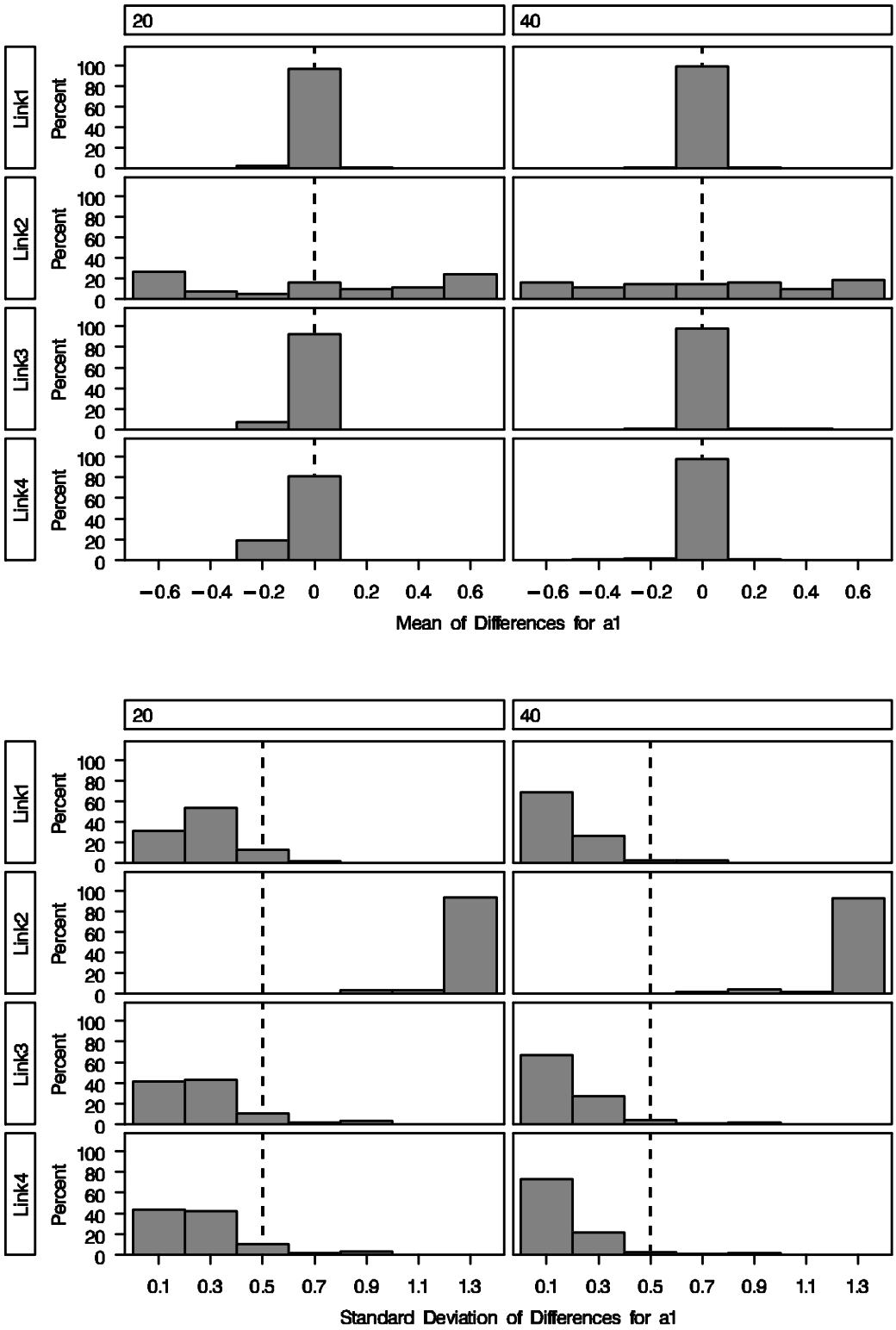


Figure 3-8. Accuracy and stability by linking method and test length for complex structure tests: N = 2000

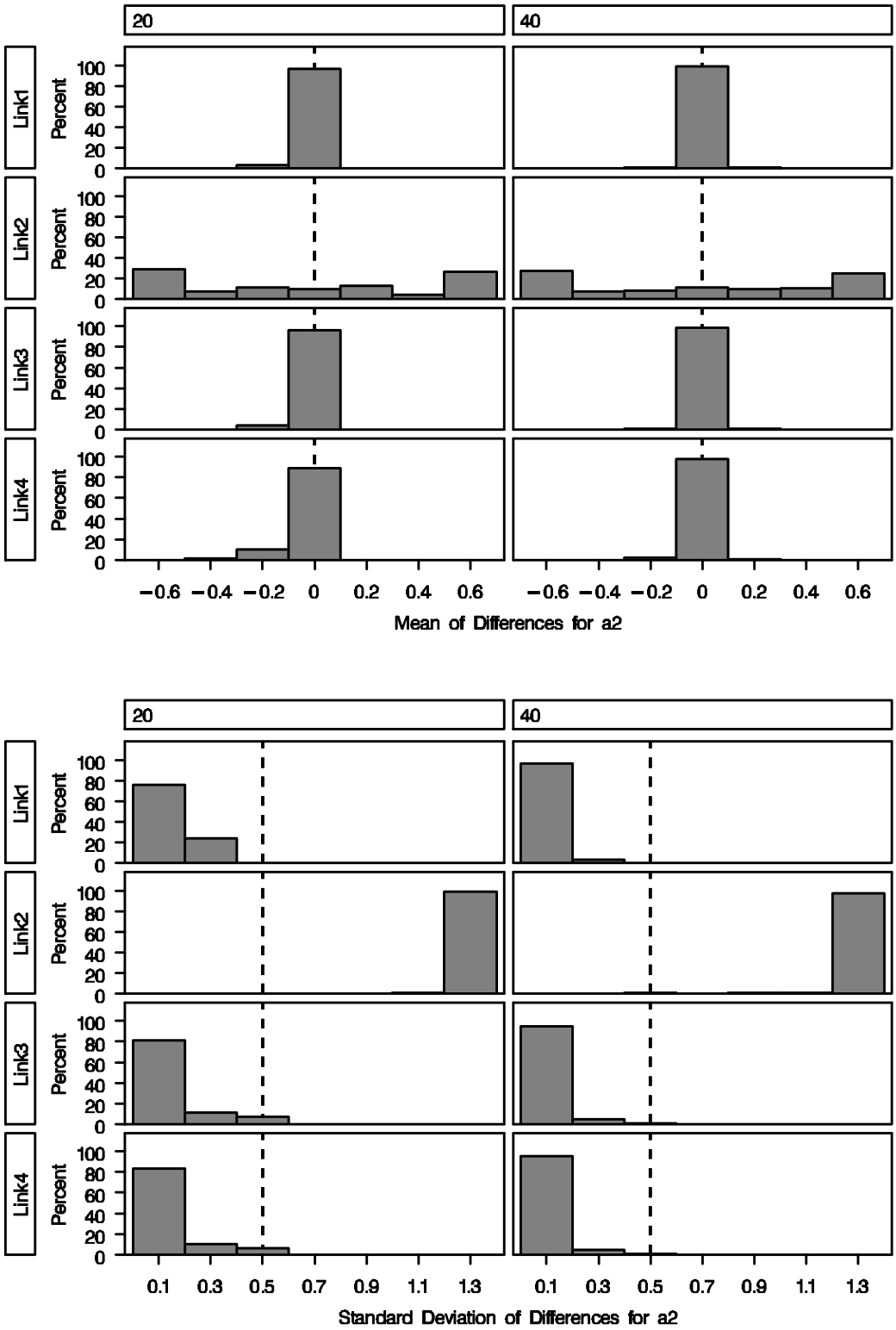


Figure 3-8. Continued

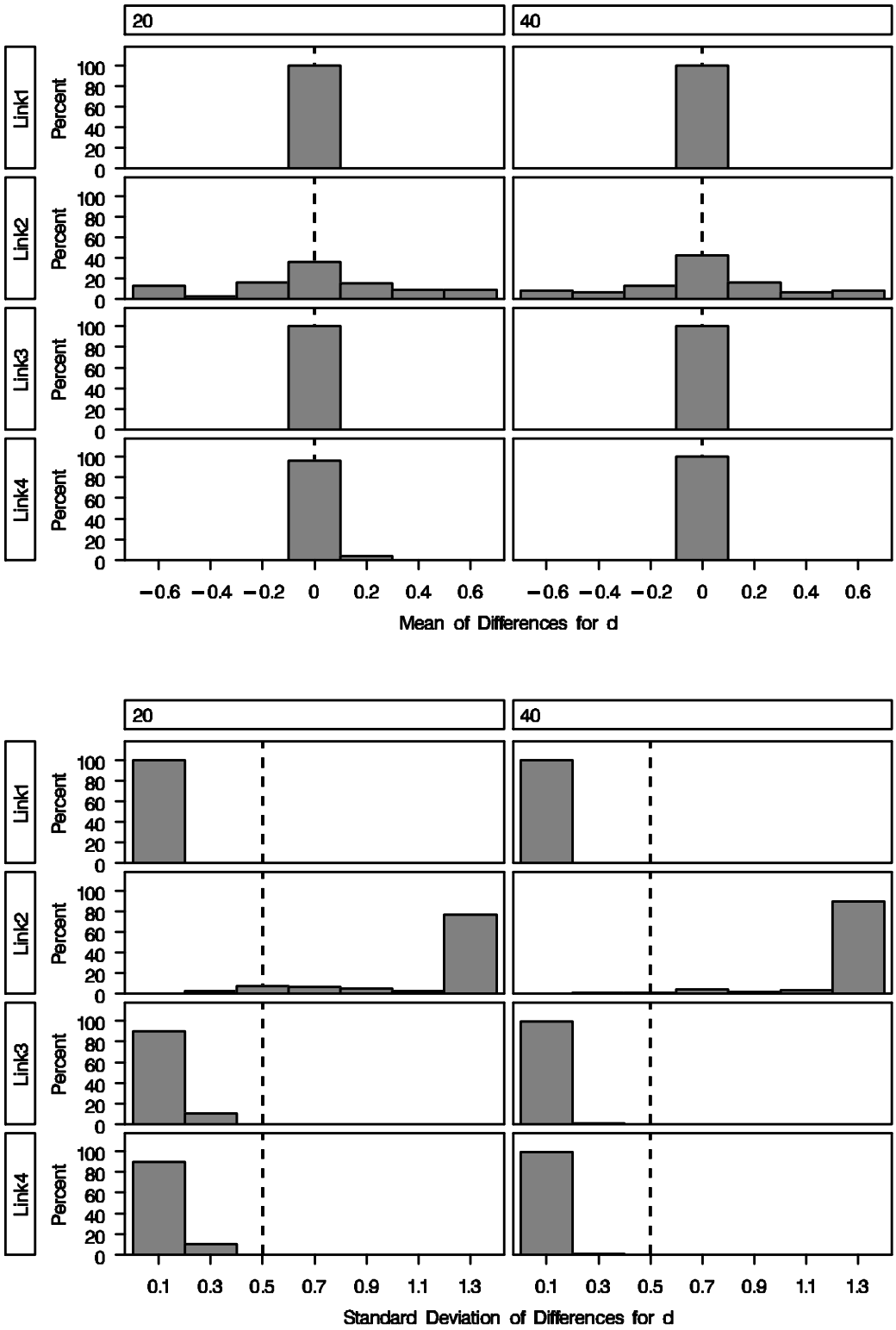


Figure 3-8. Continued

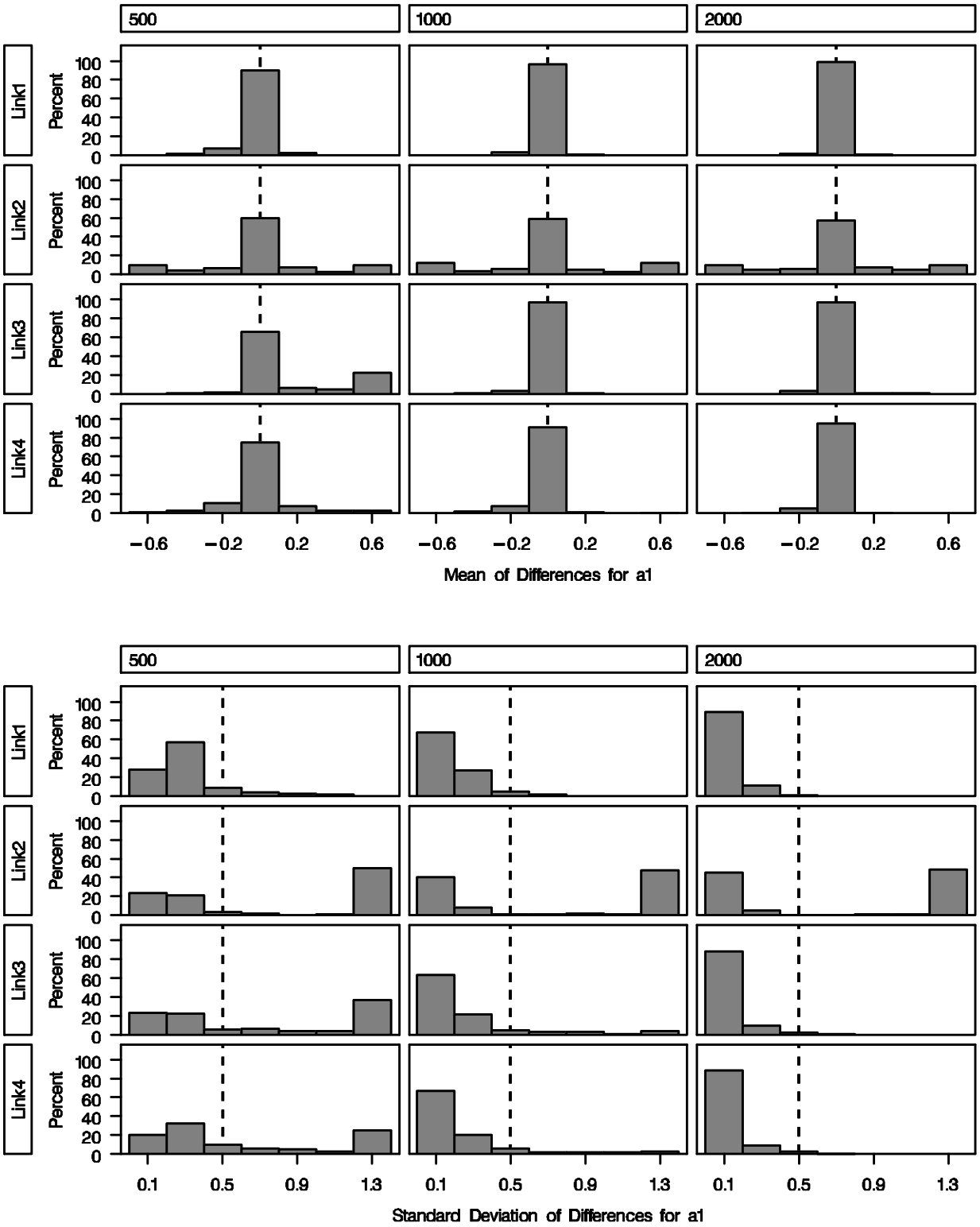


Figure 3-9. Accuracy and stability by linking method and sample size

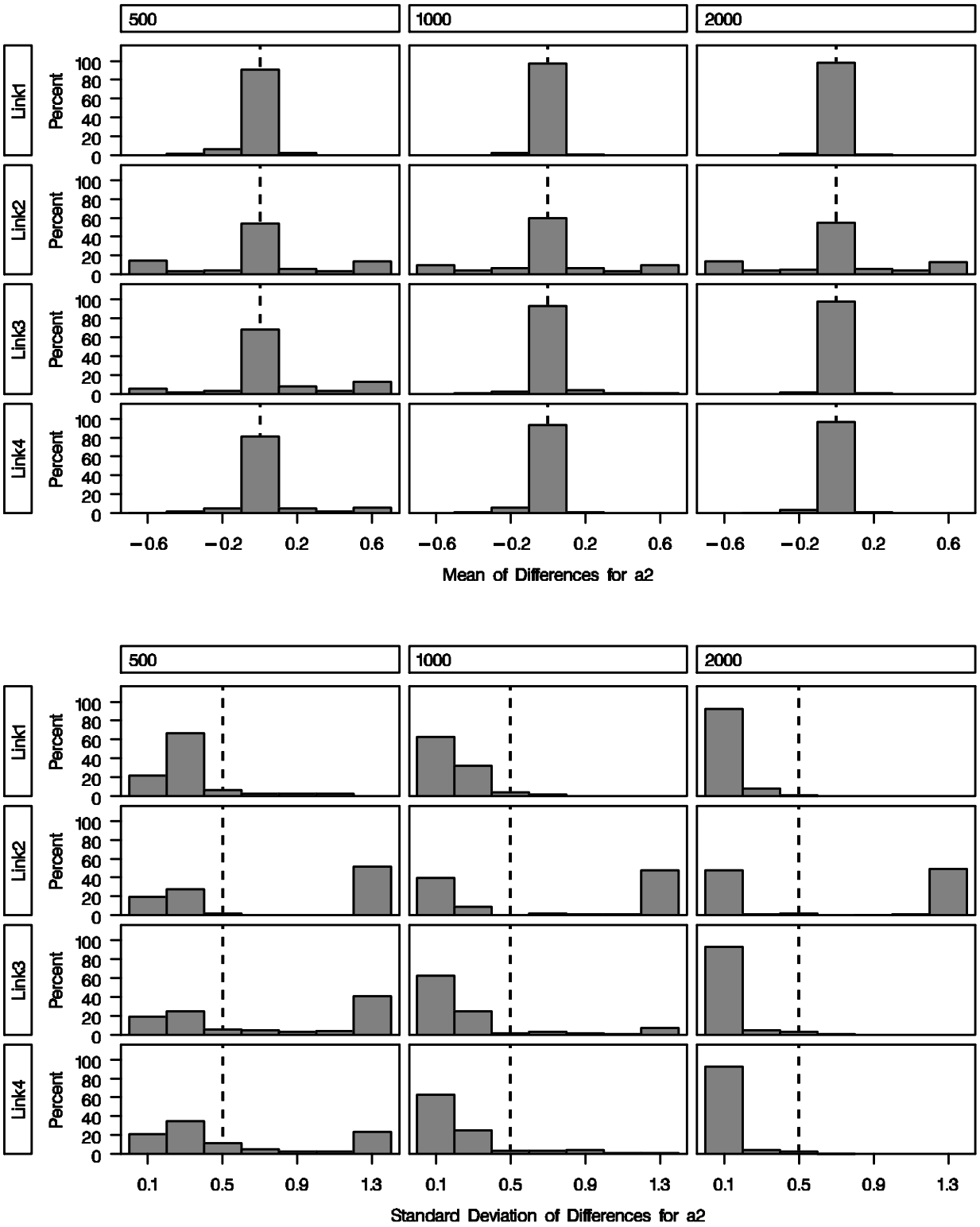
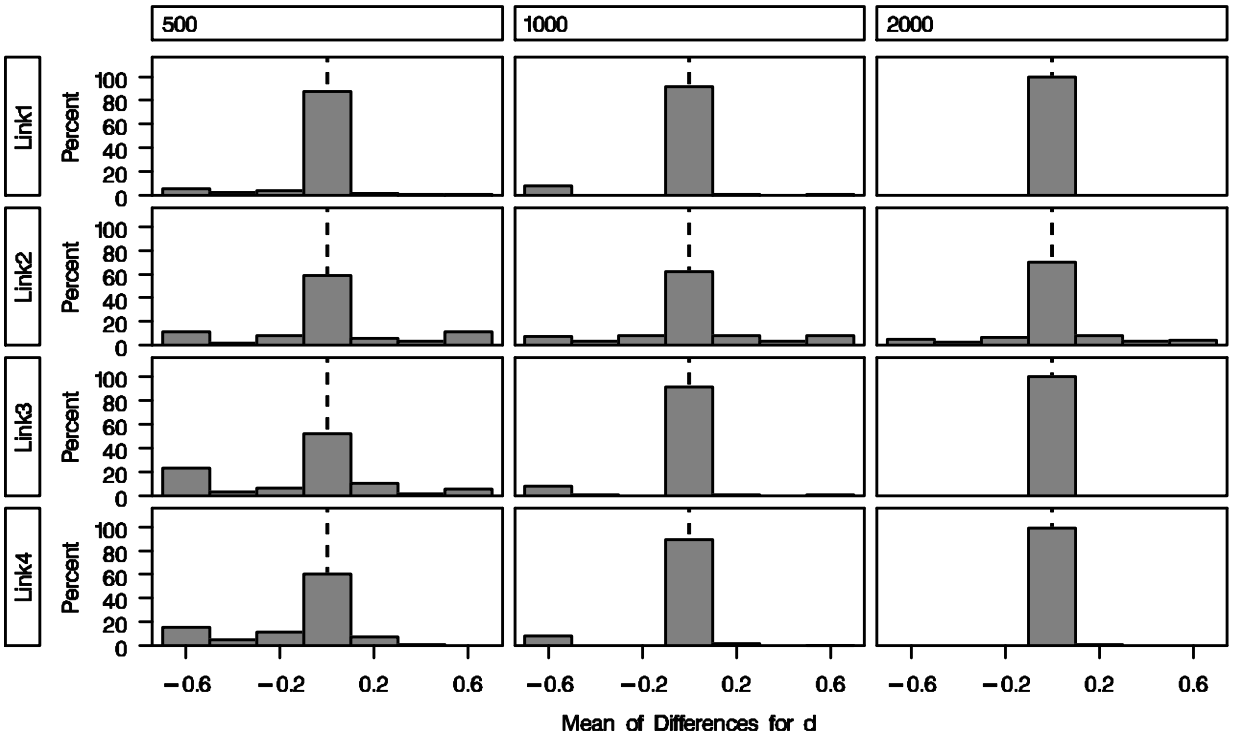


Figure 3-9. Continued



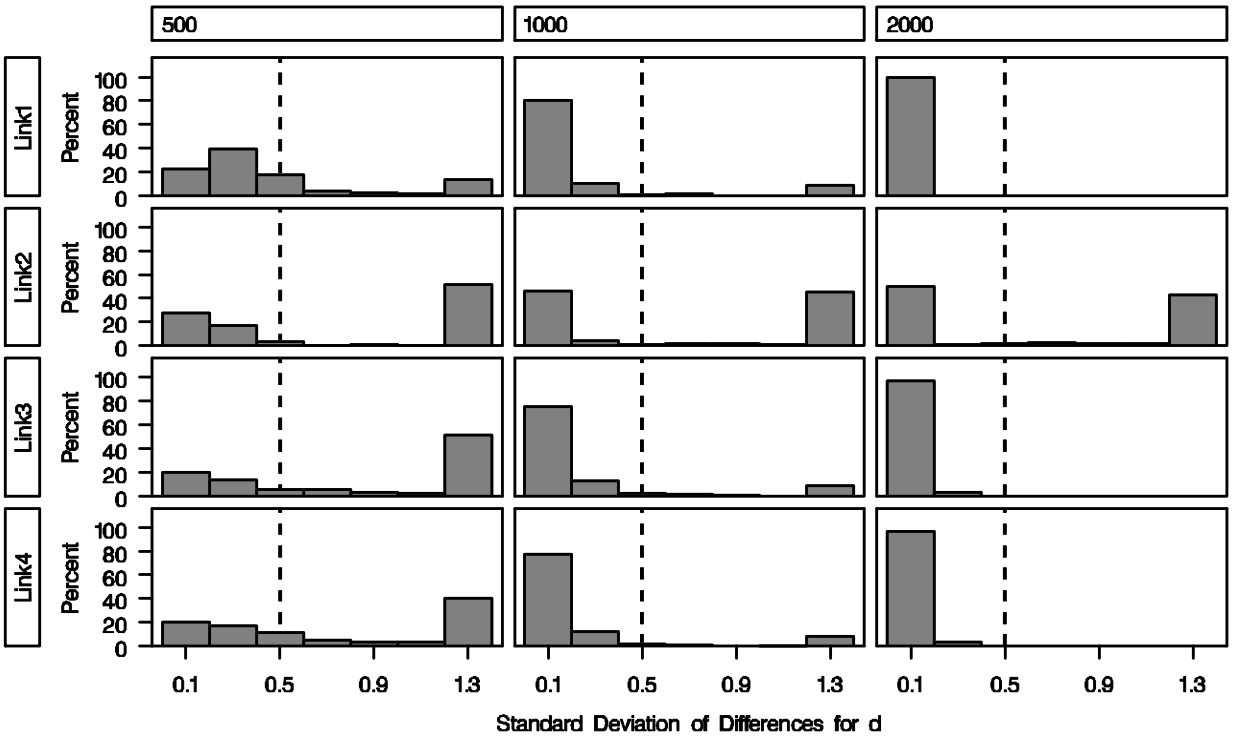


Figure 3-9. Continued

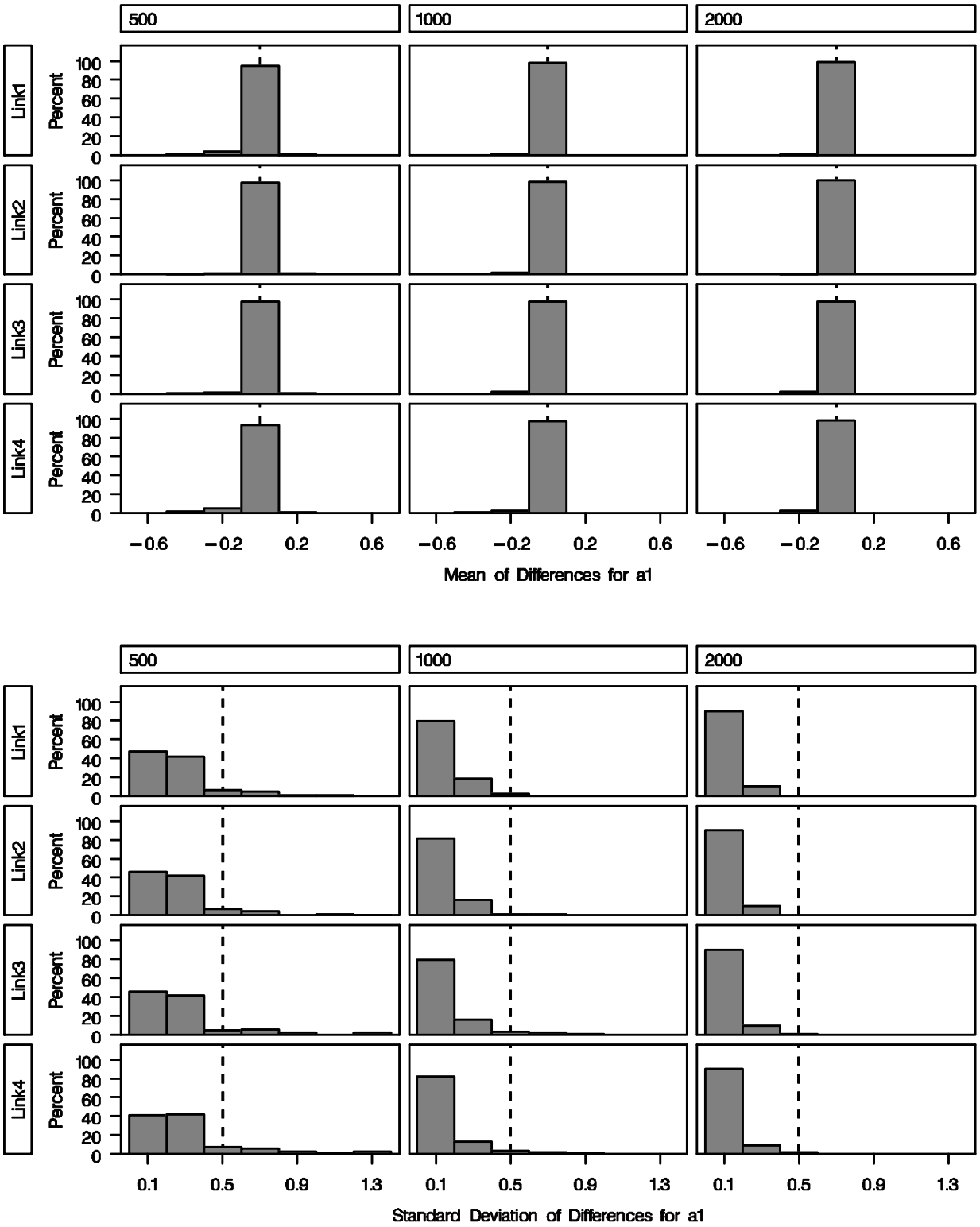


Figure 3-10. Accuracy and stability by linking method and sample size for approximate simple structure tests

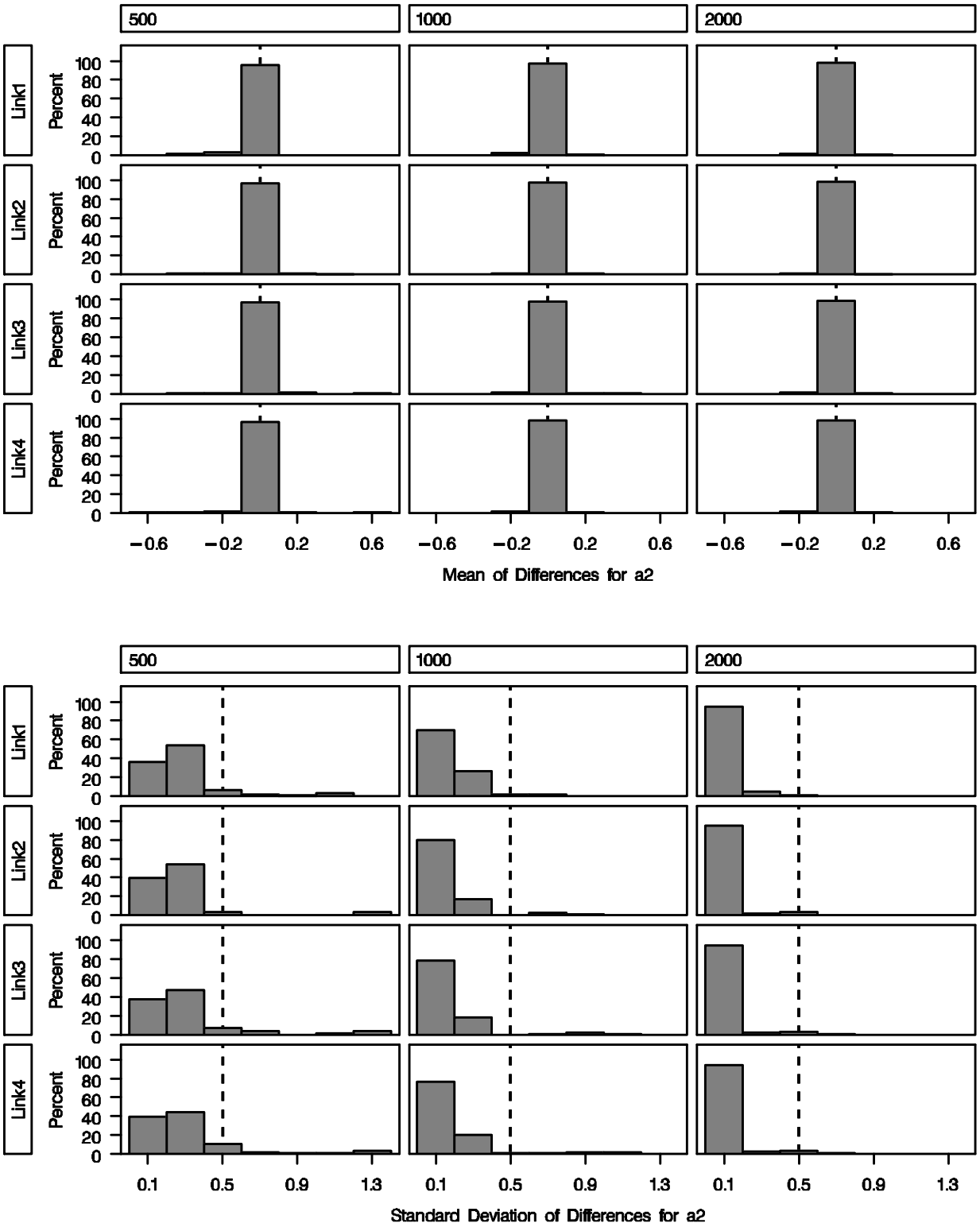


Figure 3-10. Continued

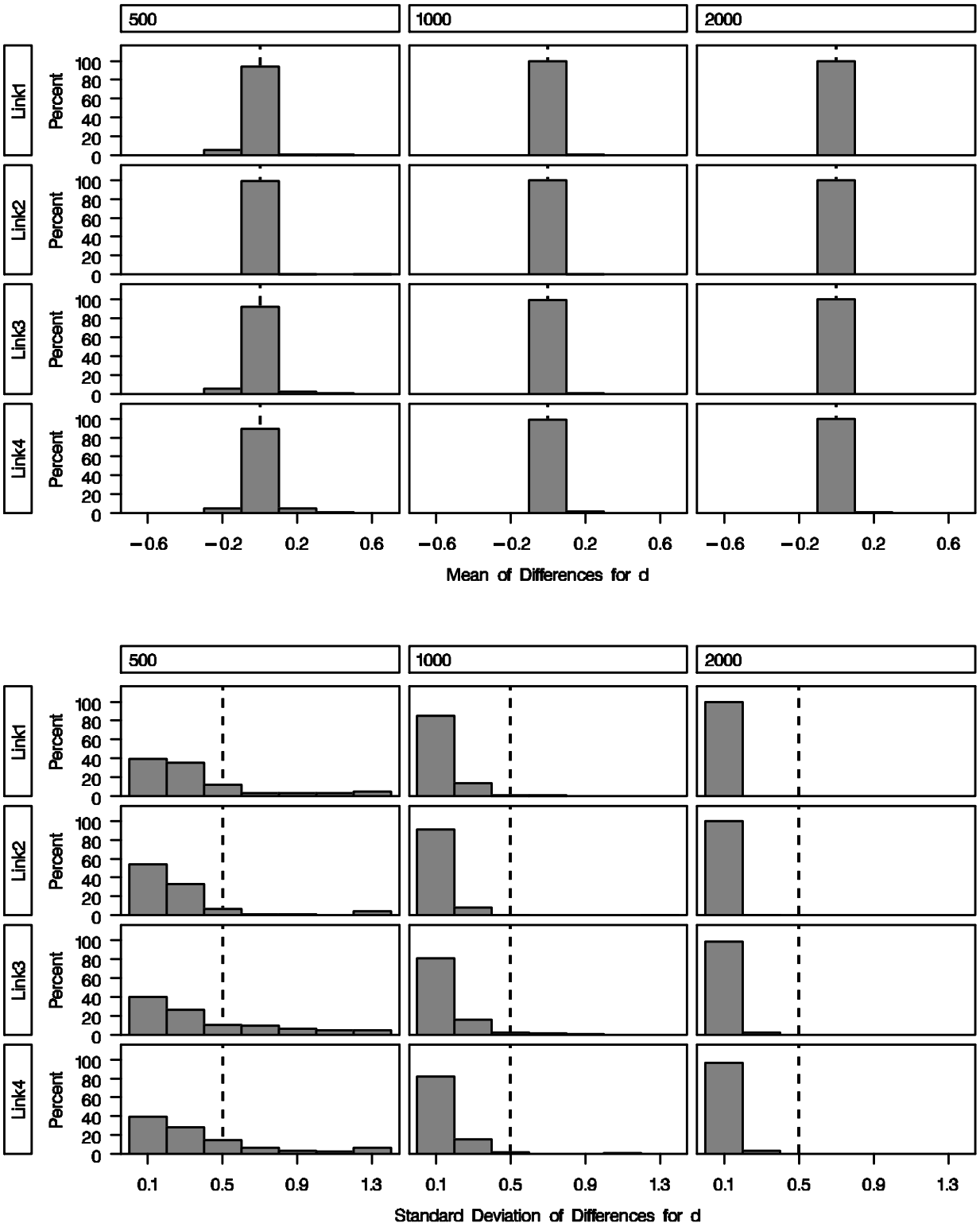


Figure 3-10. Continued

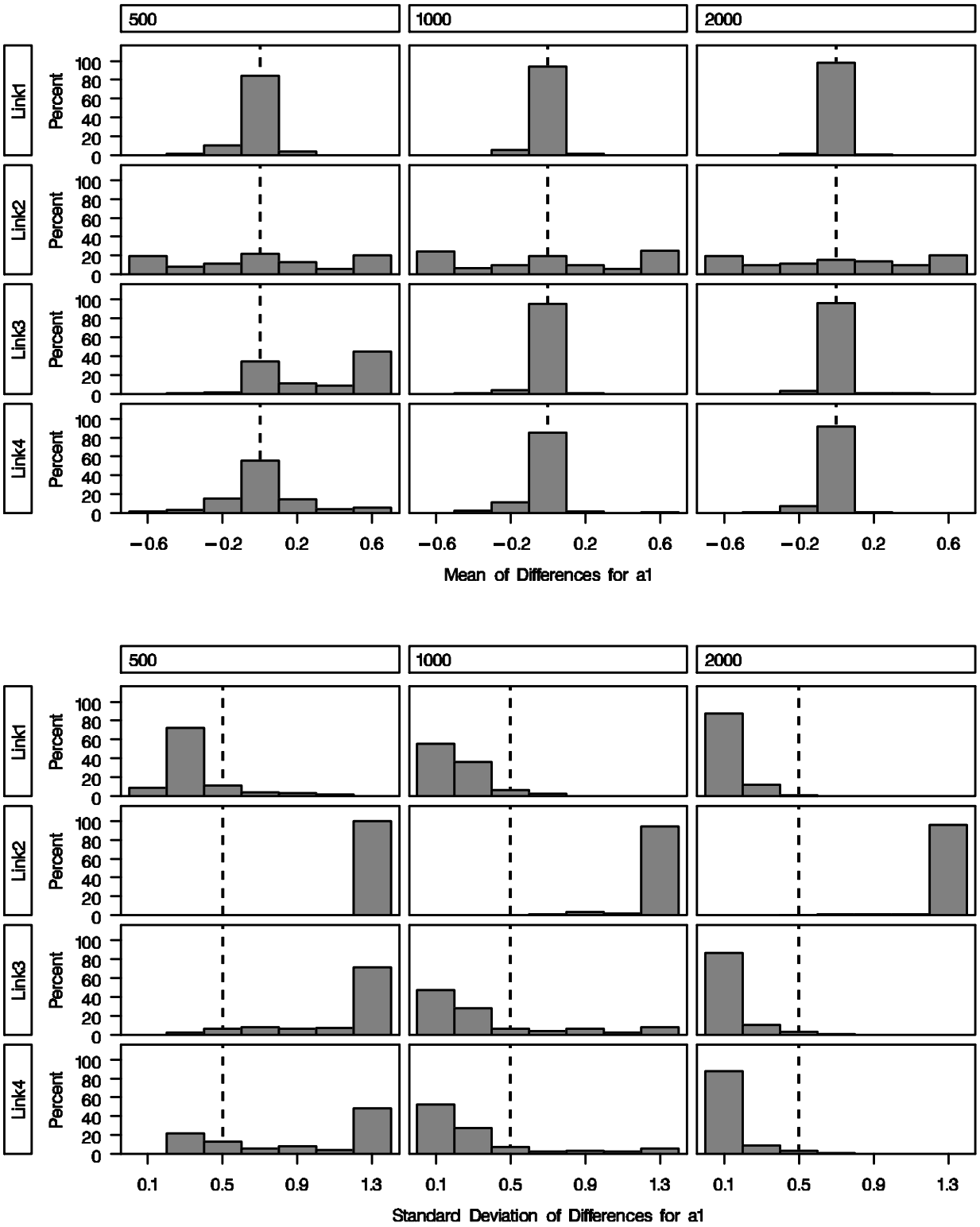


Figure 3-11. Accuracy and stability by linking method and sample size for complex structure tests

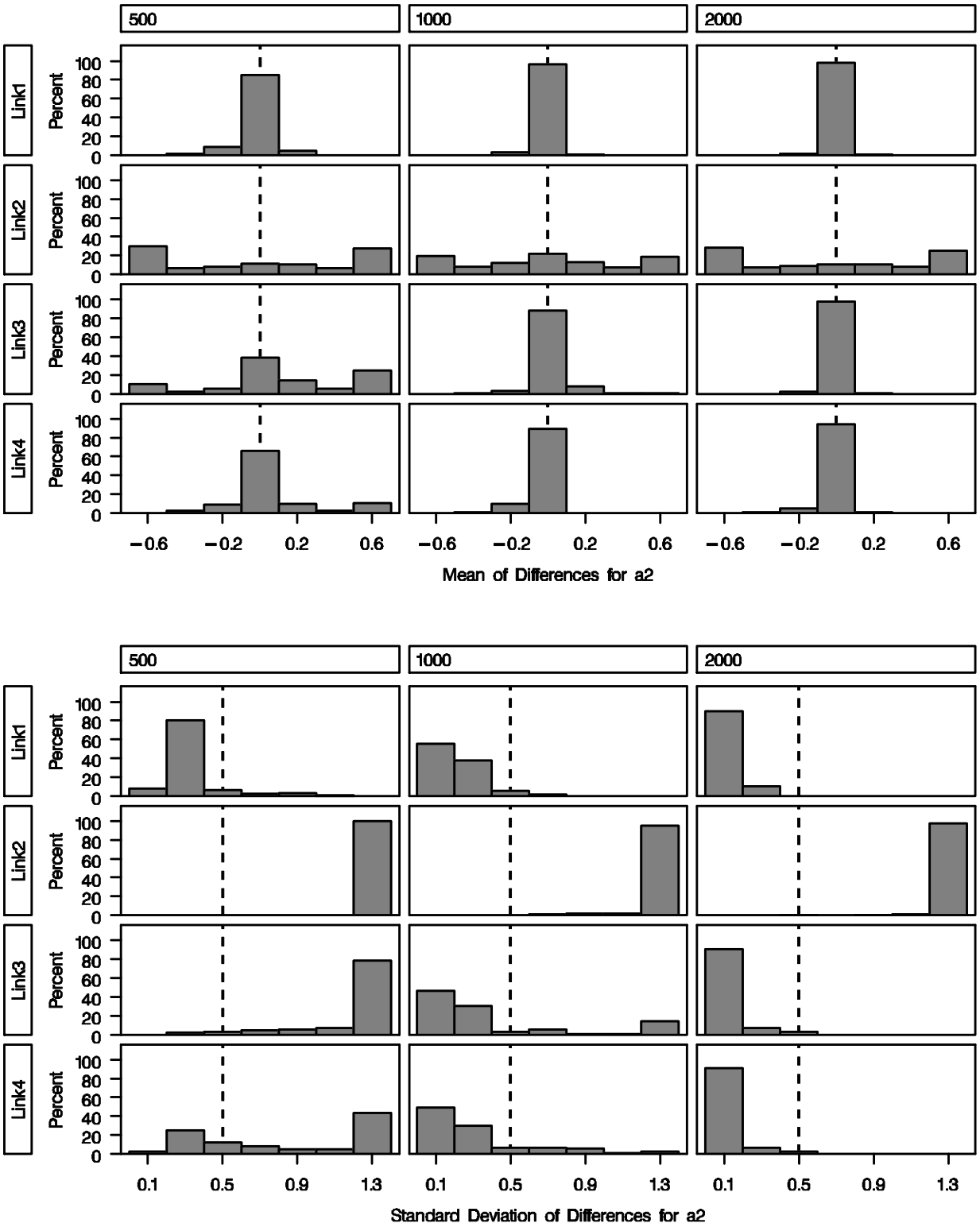


Figure 3-11. Continued

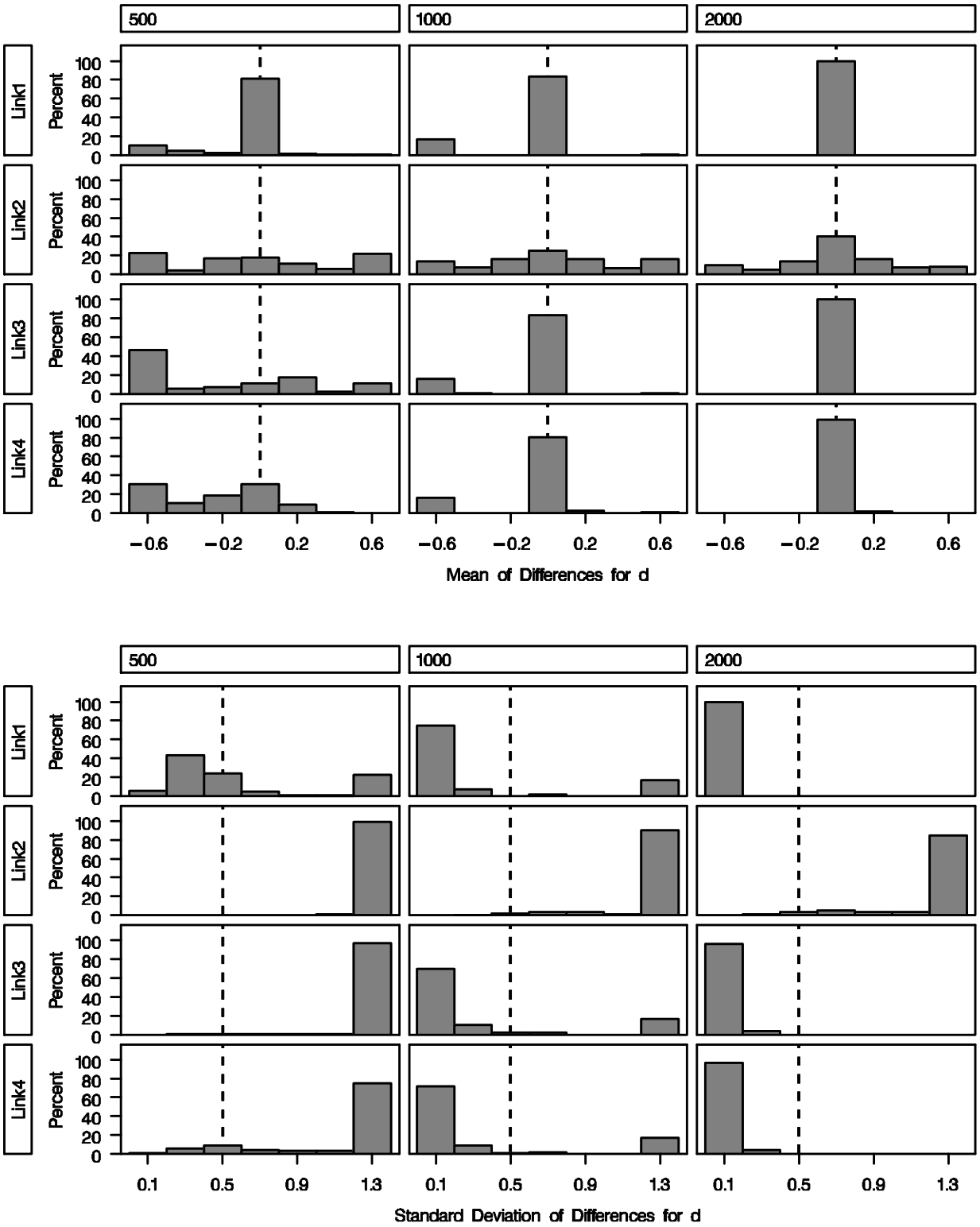


Figure 3-11. Continued

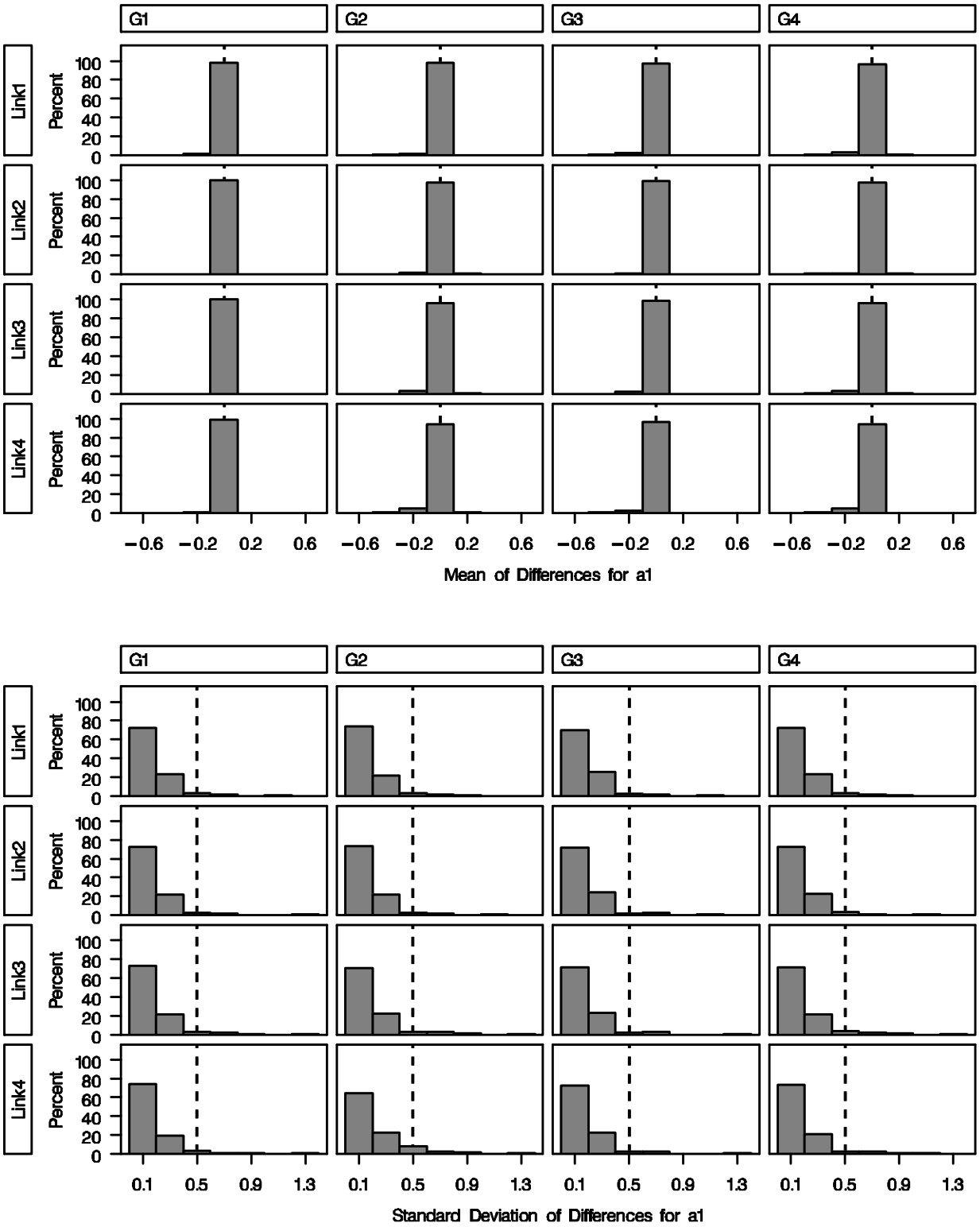


Figure 3-12. Accuracy and stability by linking method and group for approximate simple structure tests

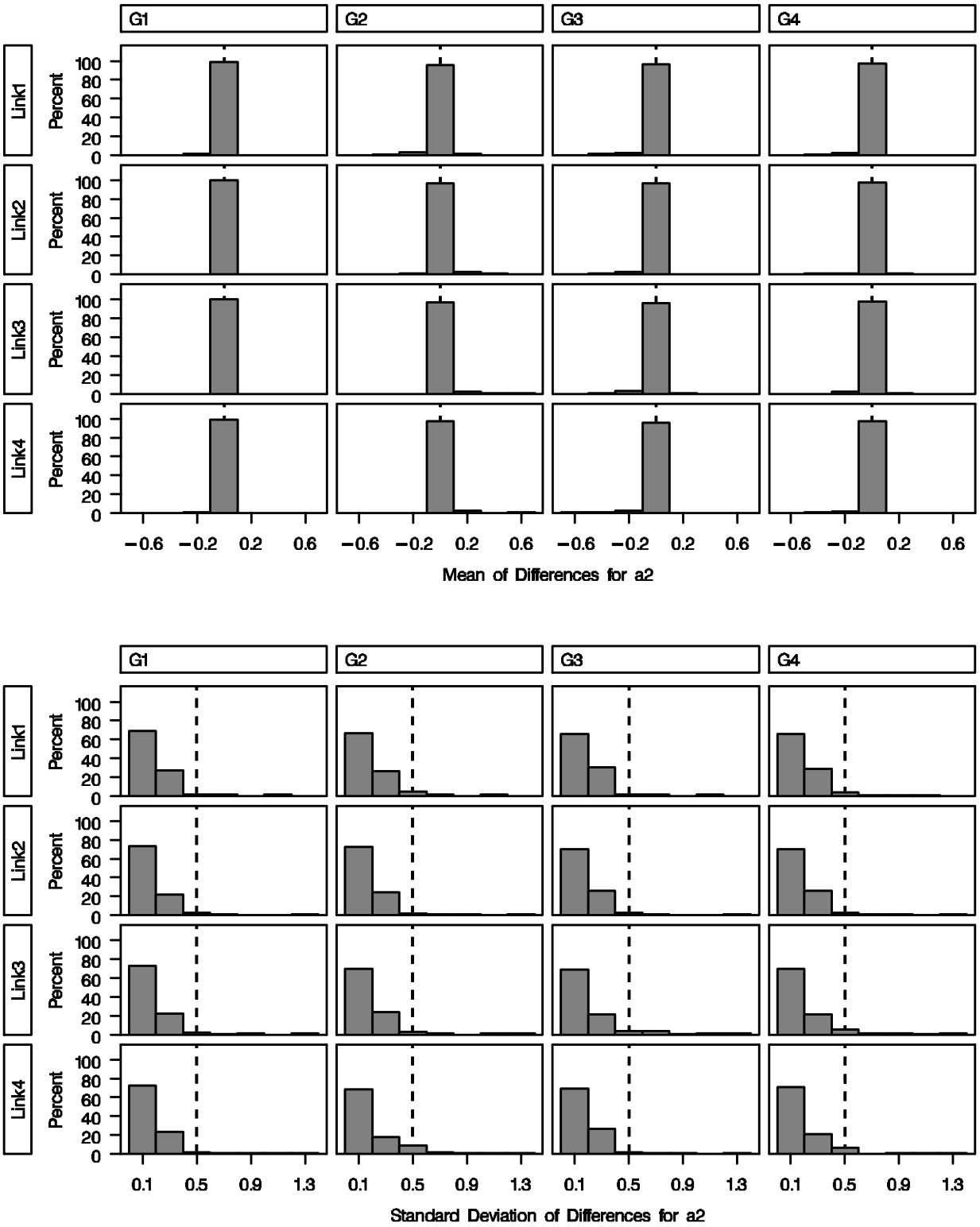
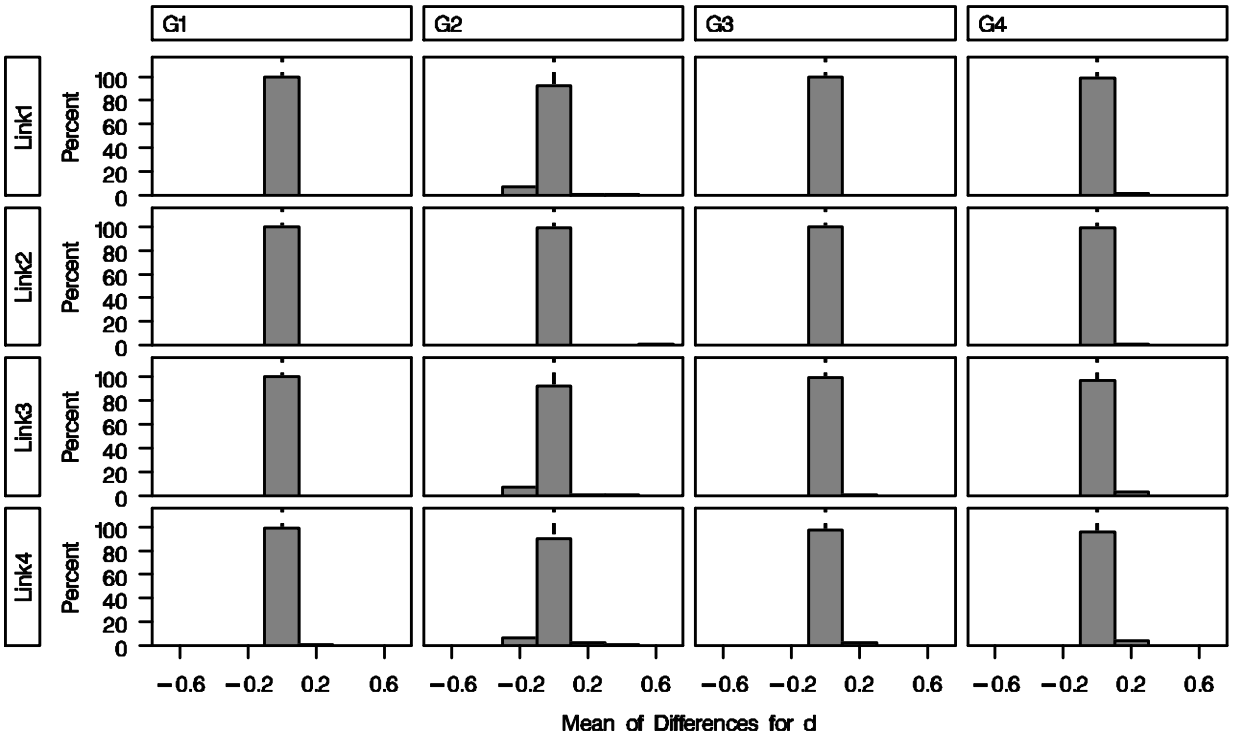


Figure 3-12. Continued



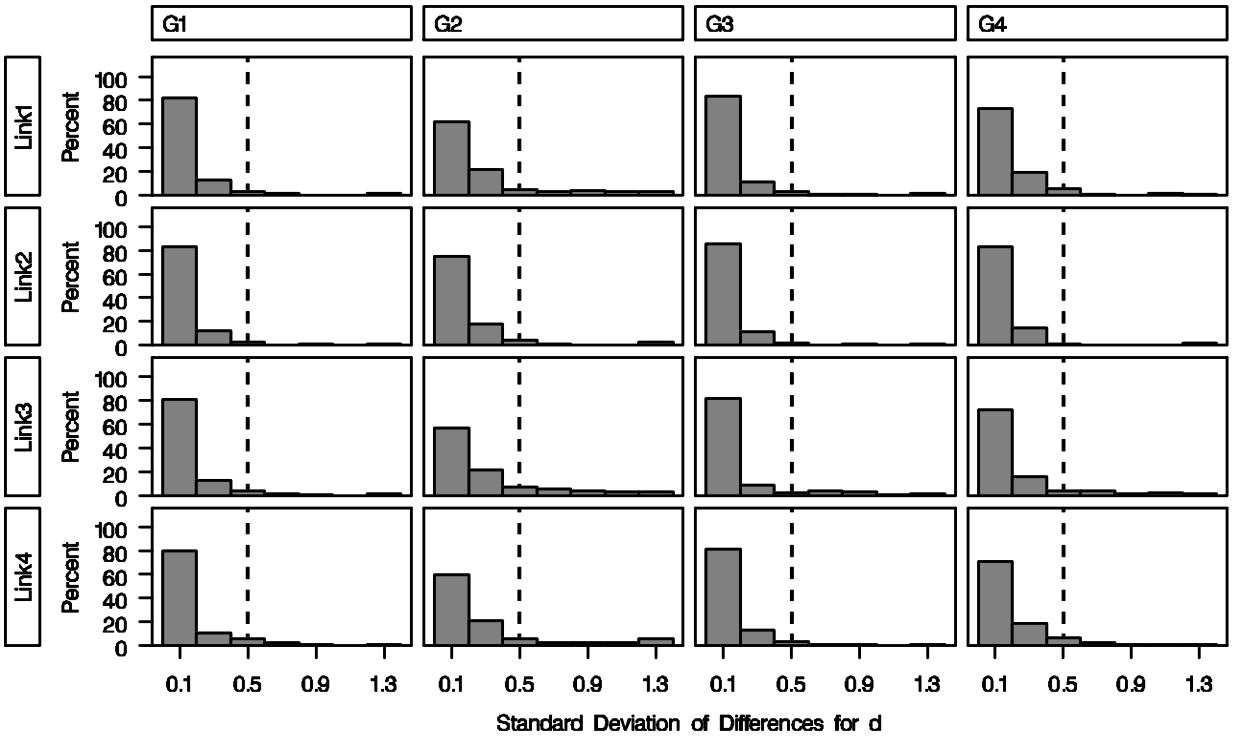


Figure 3-12. Continued

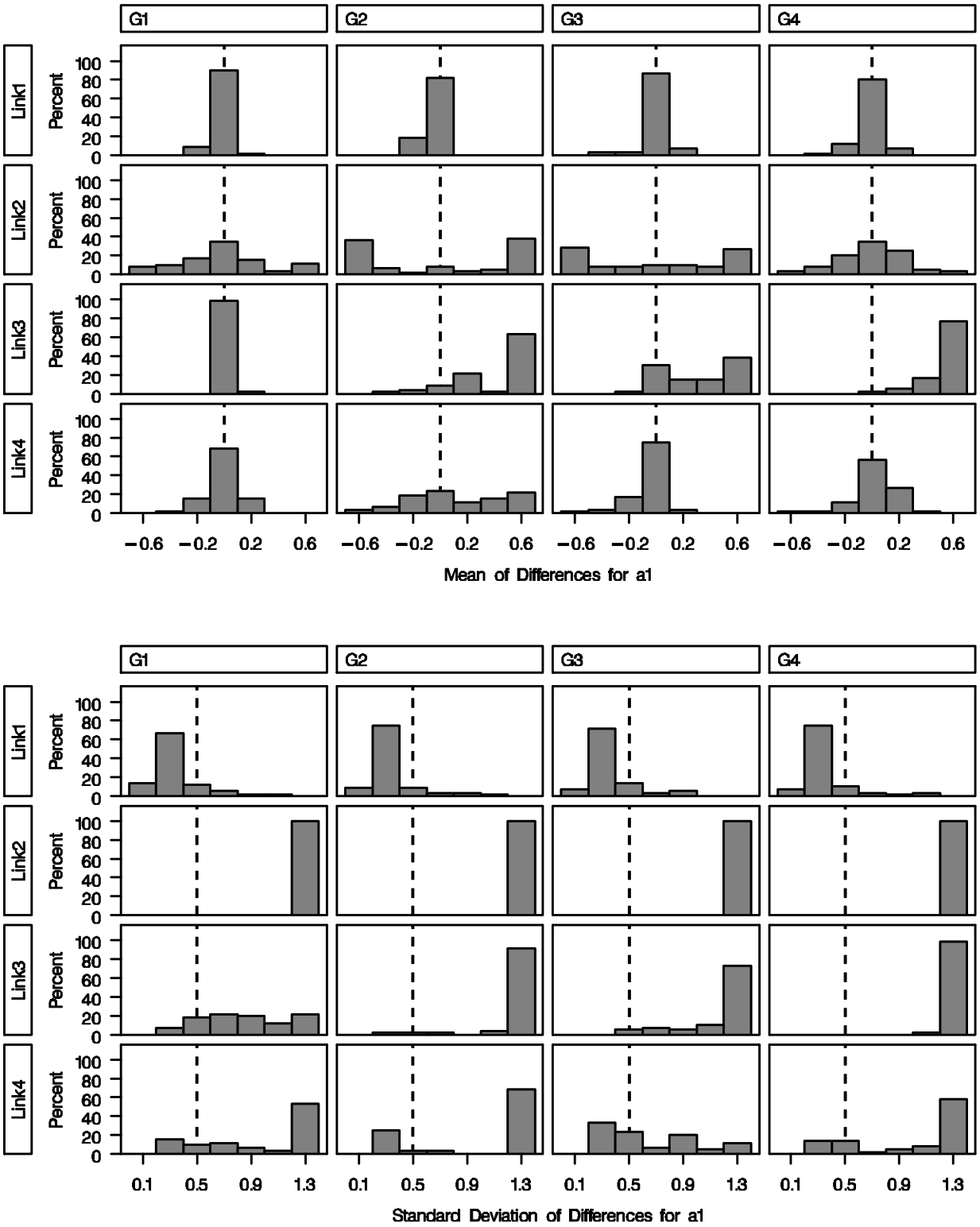
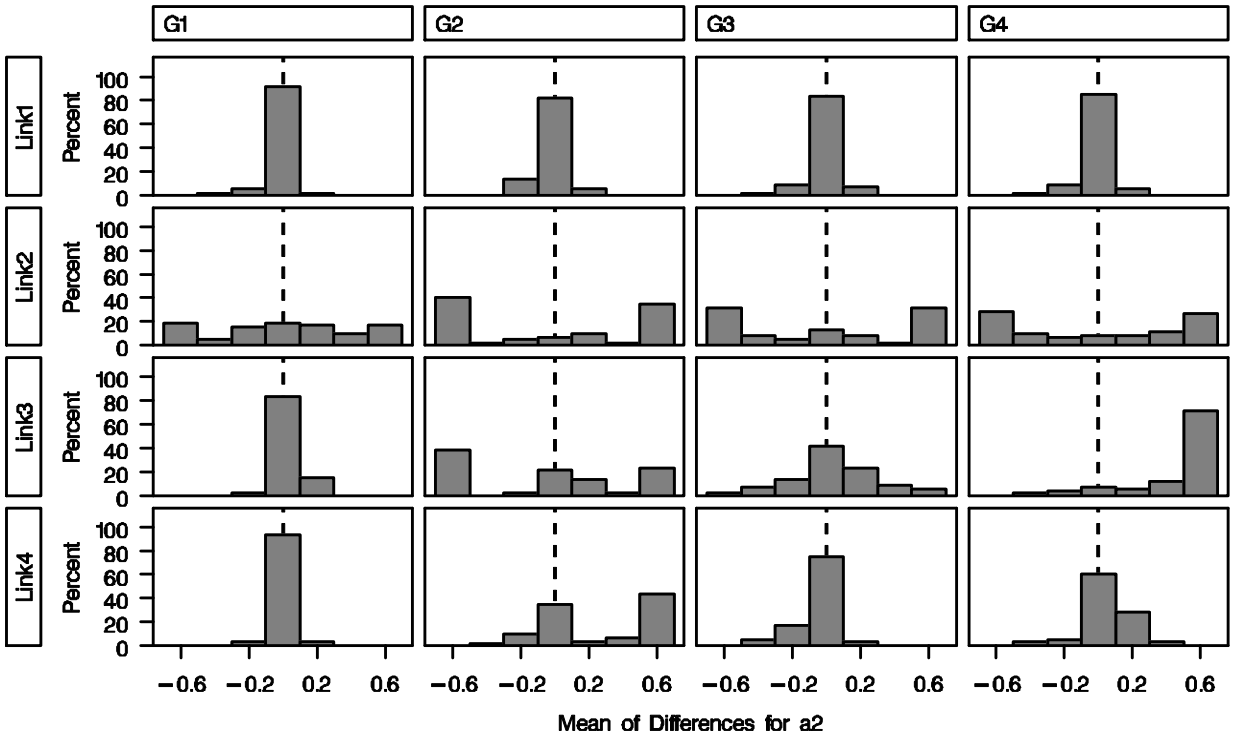


Figure 3-13. Accuracy and stability by linking method and group for complex structure tests: N = 500



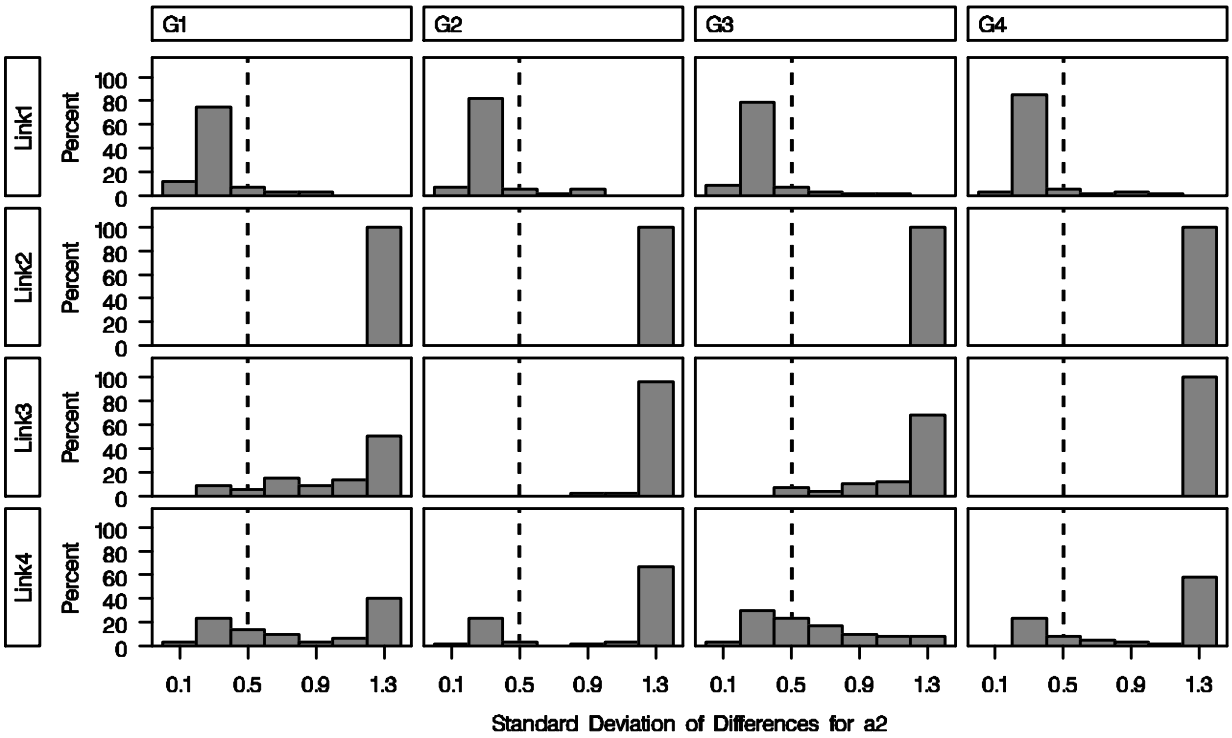


Figure 3-13. Continued

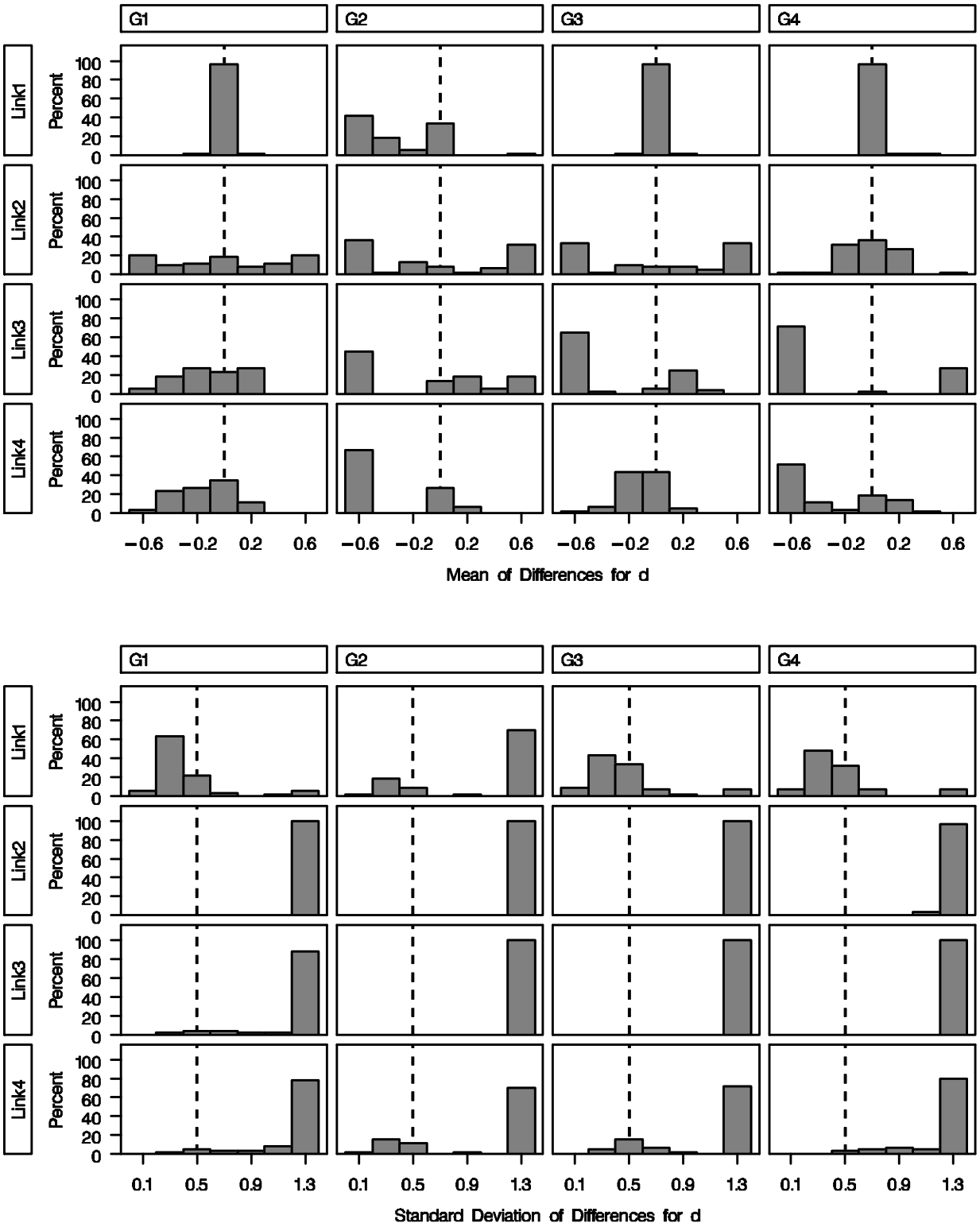


Figure 3-13. Continued

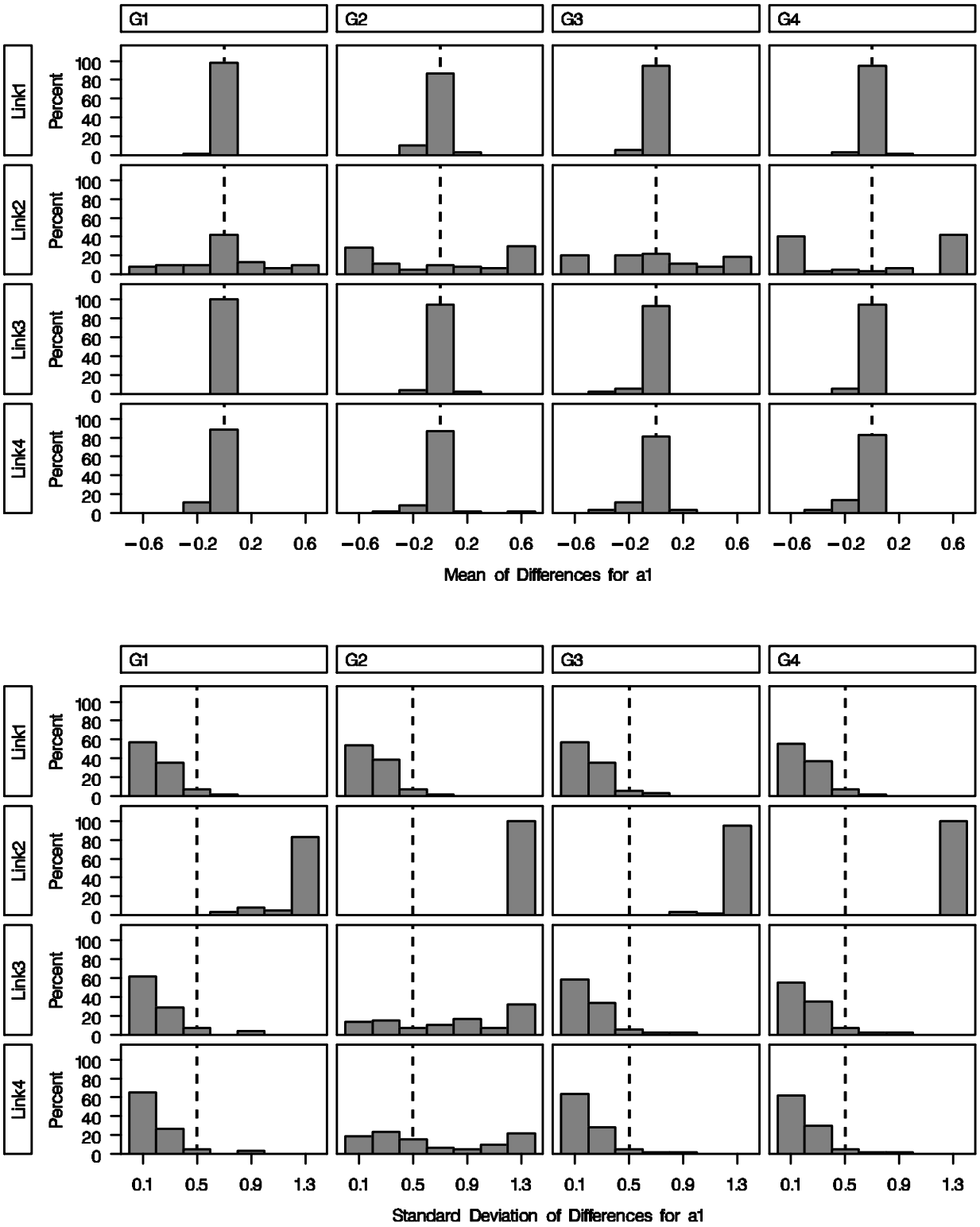


Figure 3-14. Accuracy and stability by linking method and group for complex structure tests: N = 1000

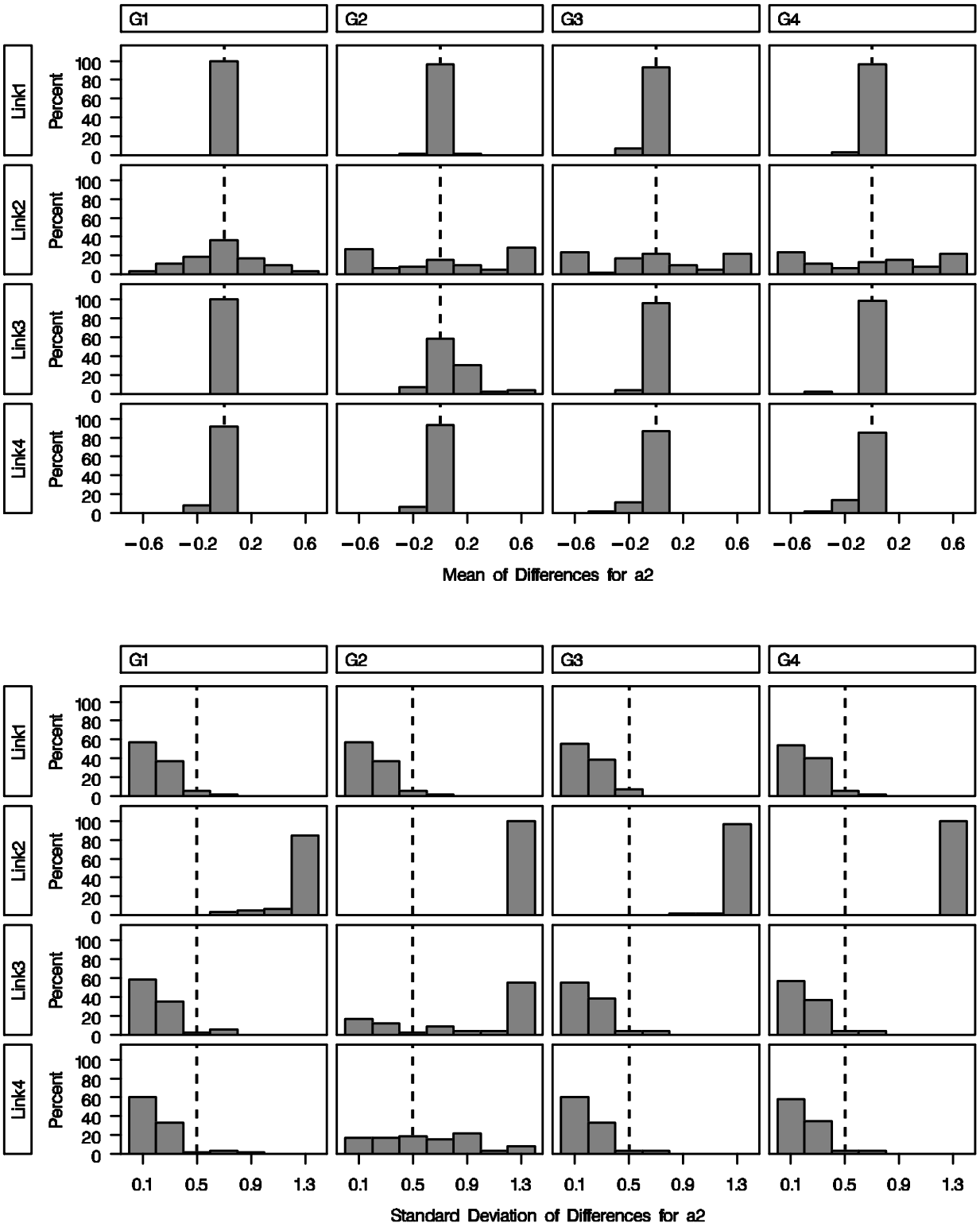


Figure 3-14. Continued

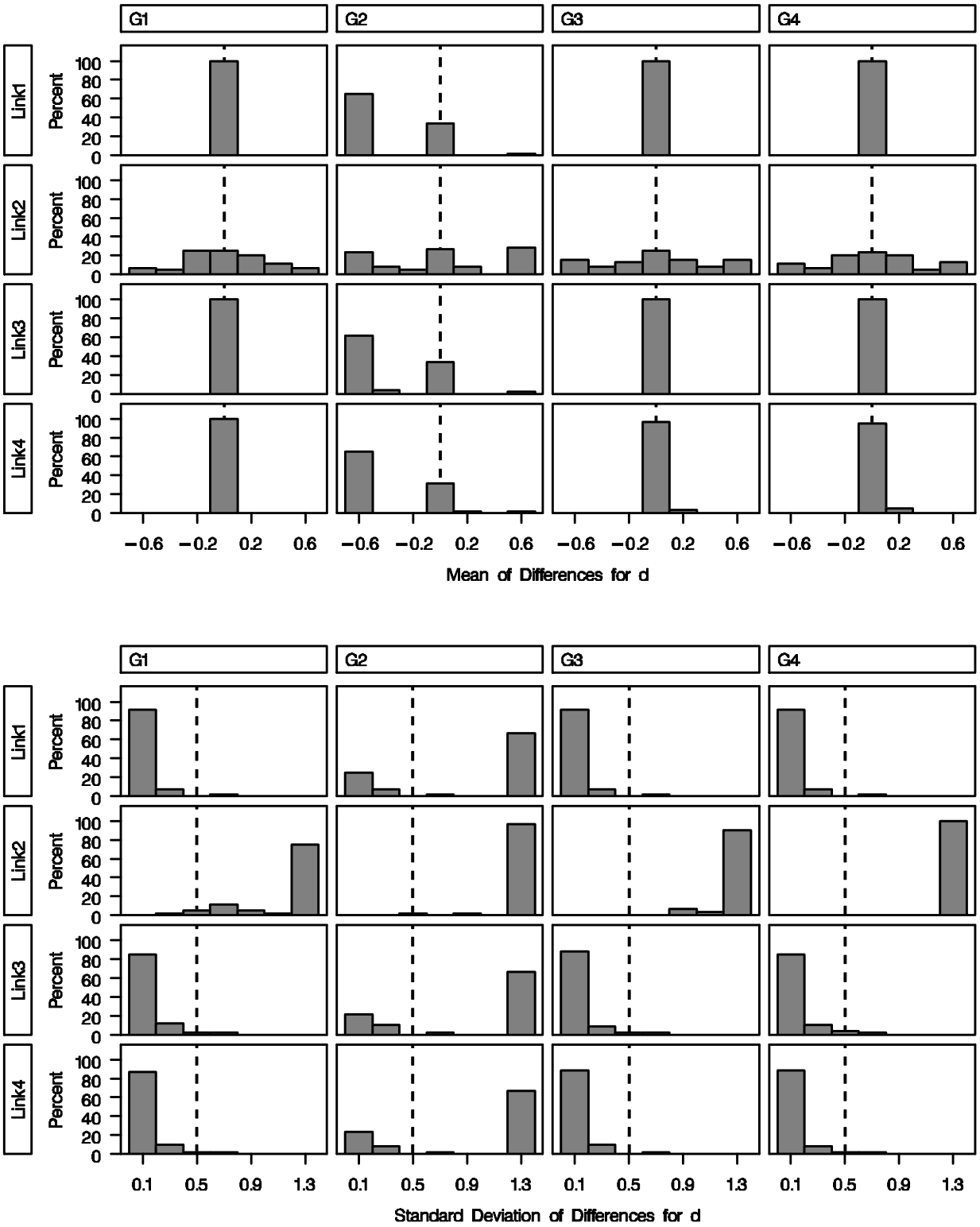


Figure 3-14. Continued

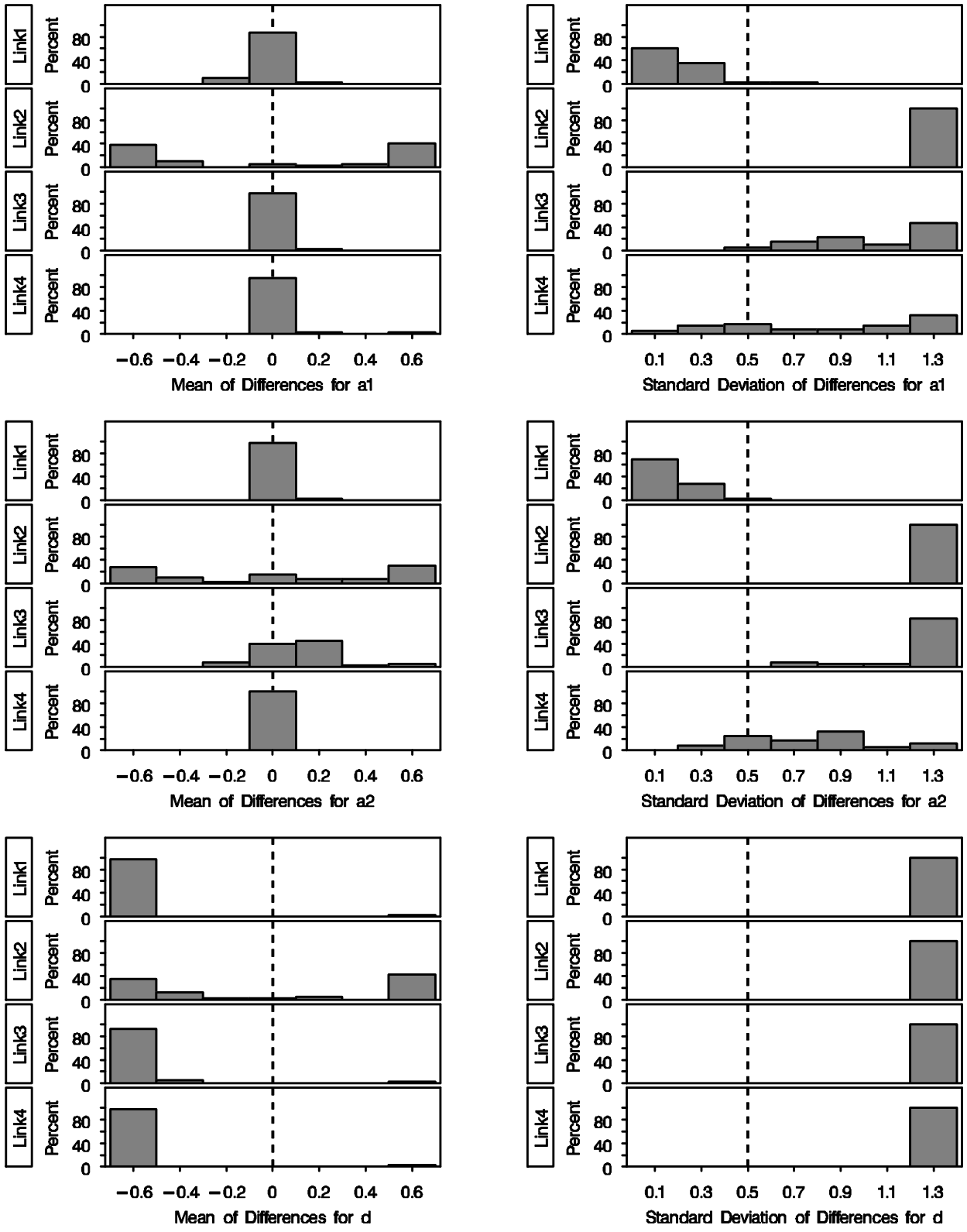


Figure 3-15. Accuracy and stability for different linking methods: COM, n=40, N=1000, G2

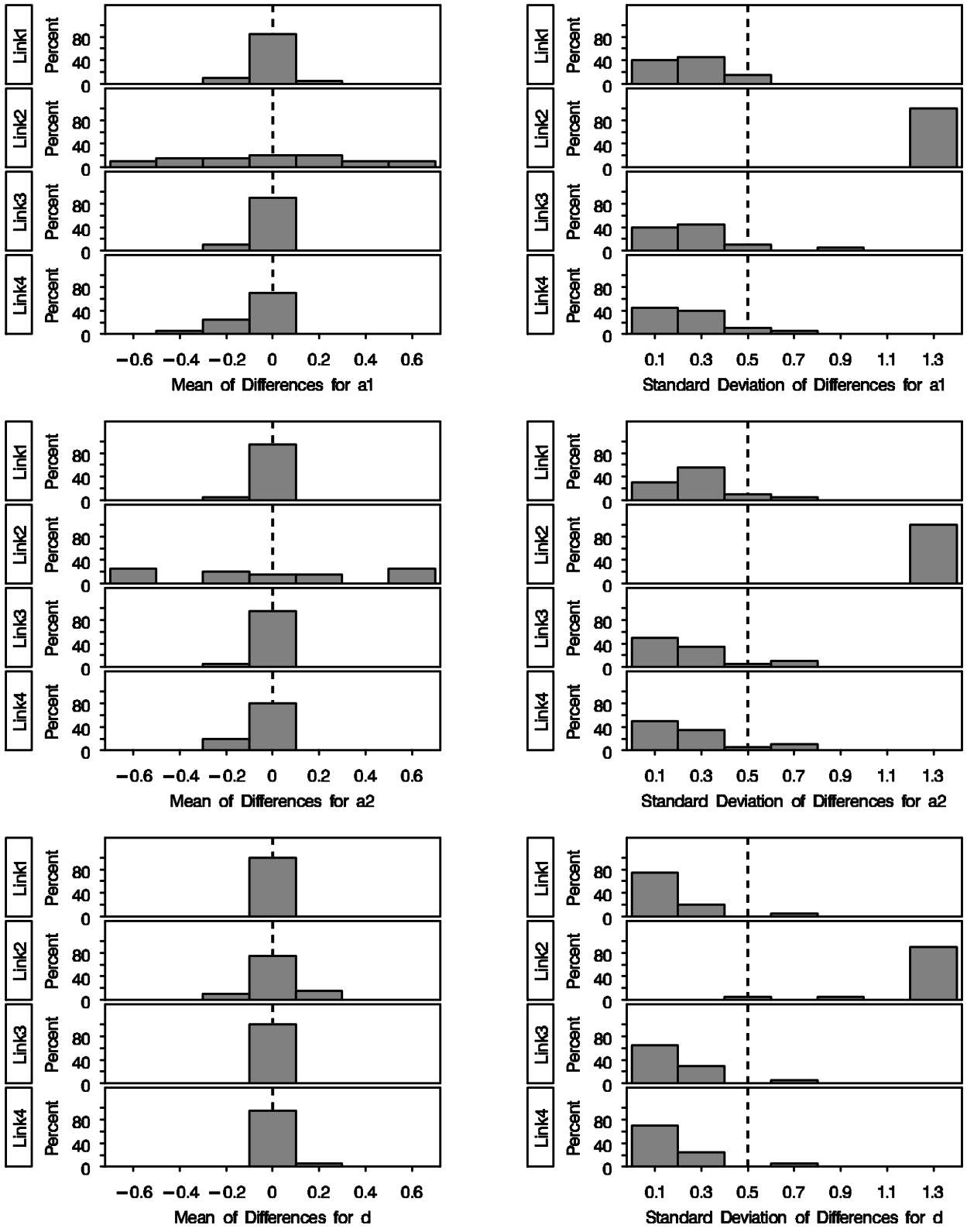


Figure 3-16. Accuracy and stability for different linking methods: COM, n=20, N=1000, G2

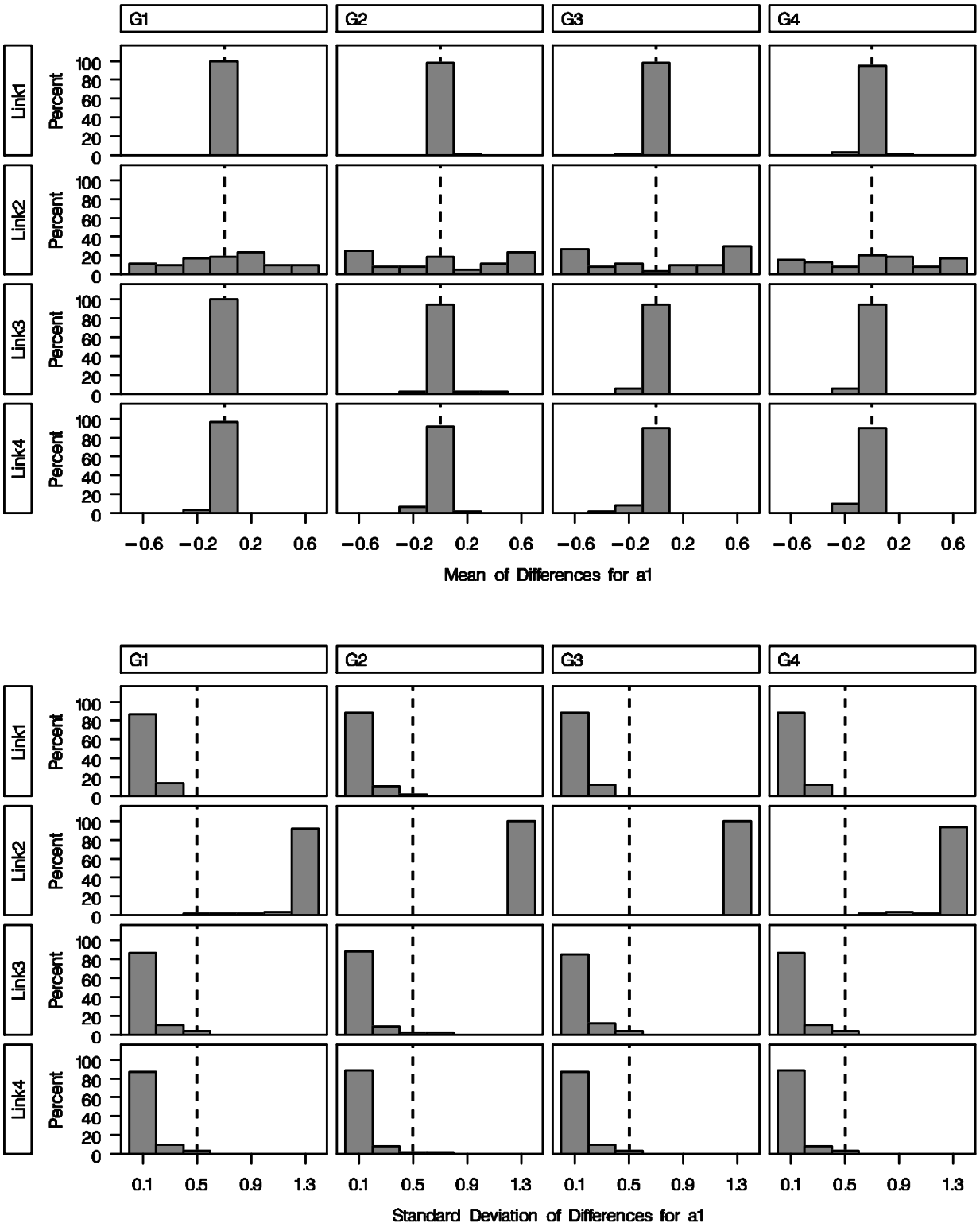
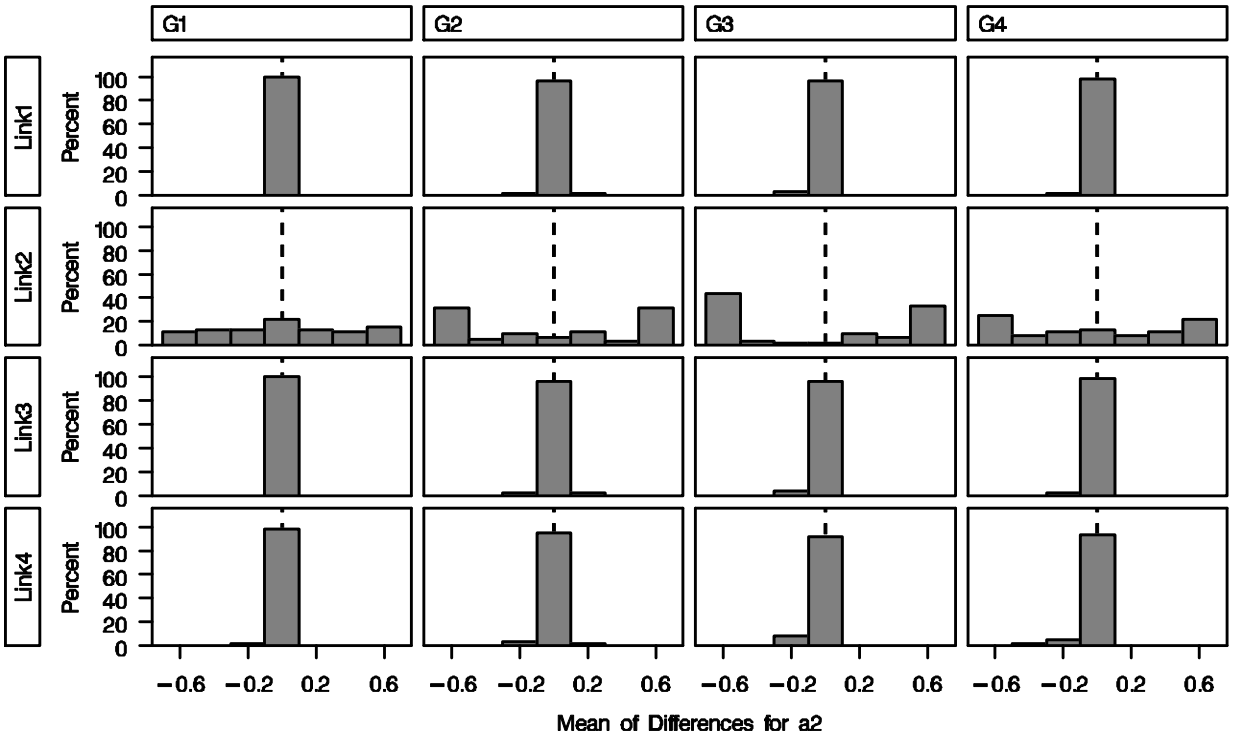


Figure 3-17. Accuracy and stability by linking method and group for complex structure tests: N = 2000



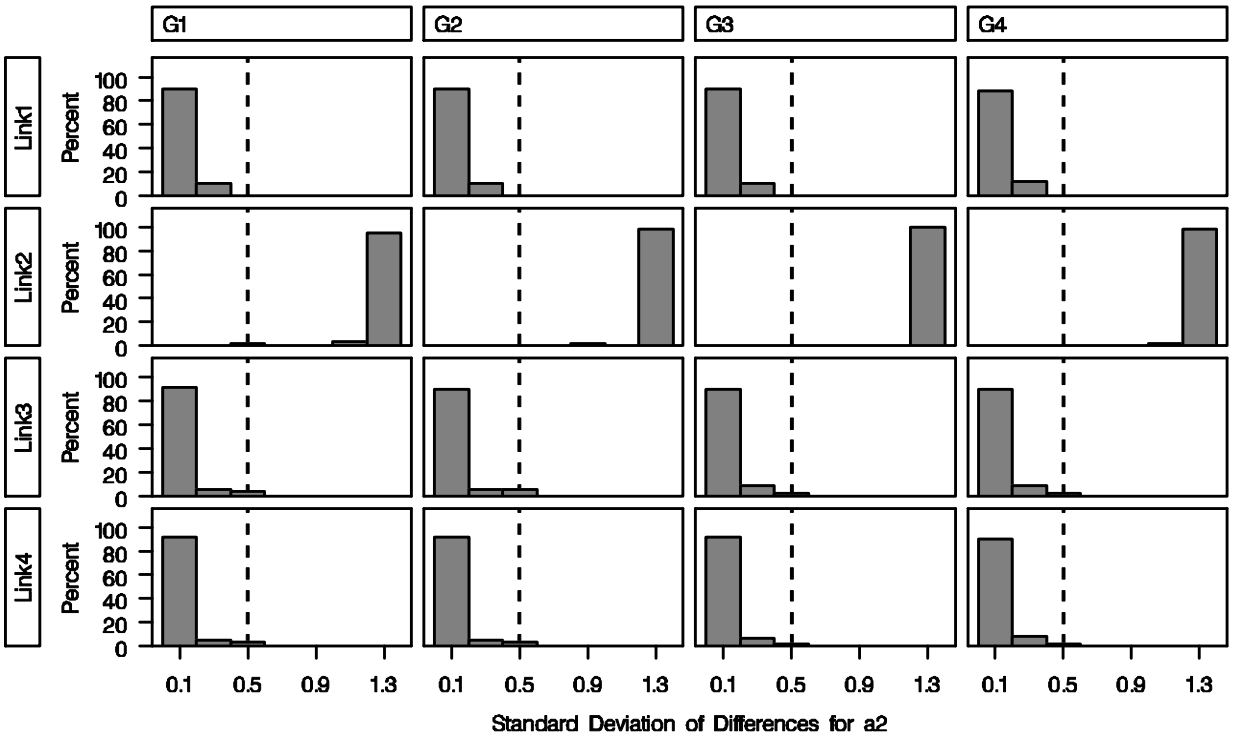


Figure 3-17. Continued

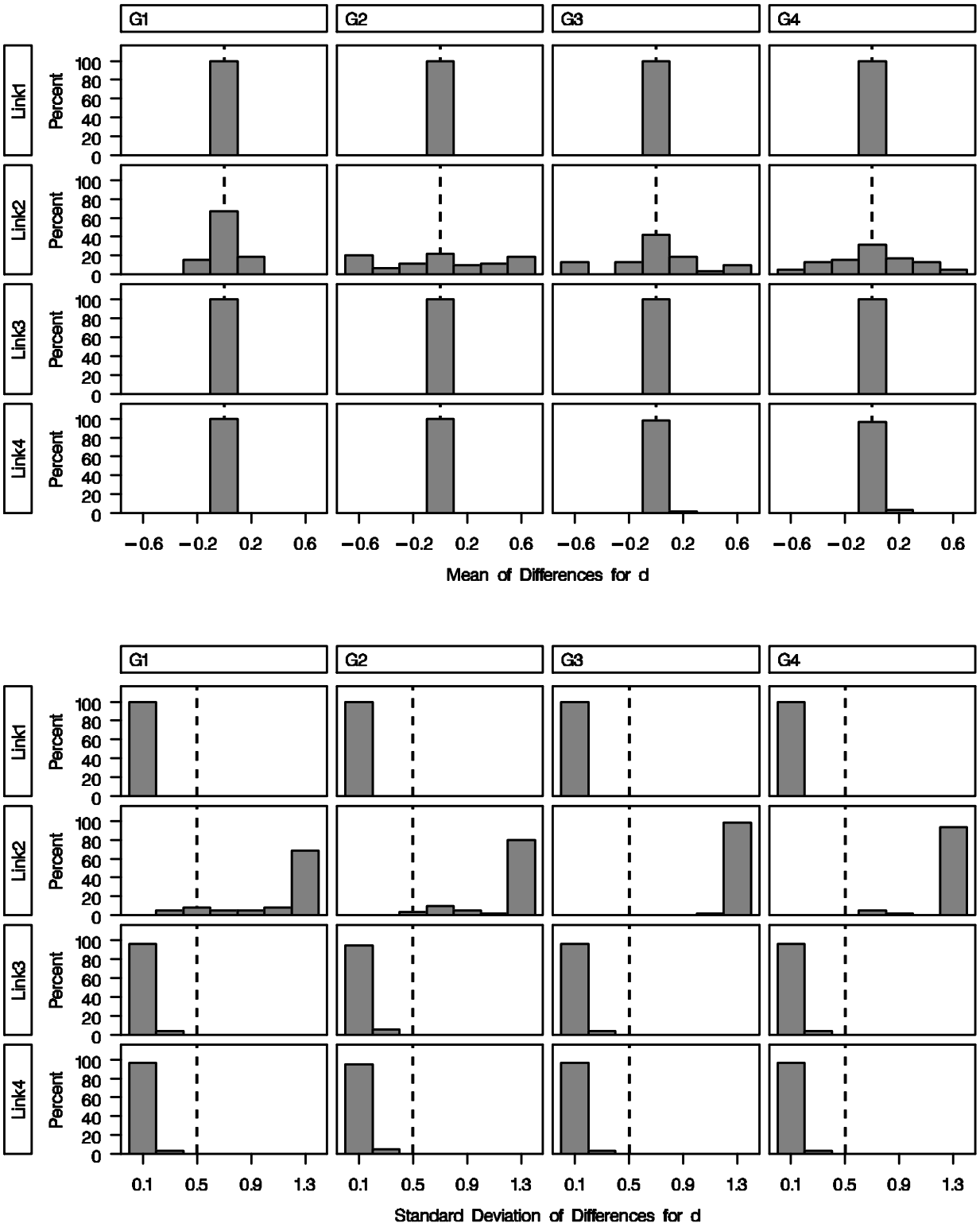


Figure 3-17. Continued

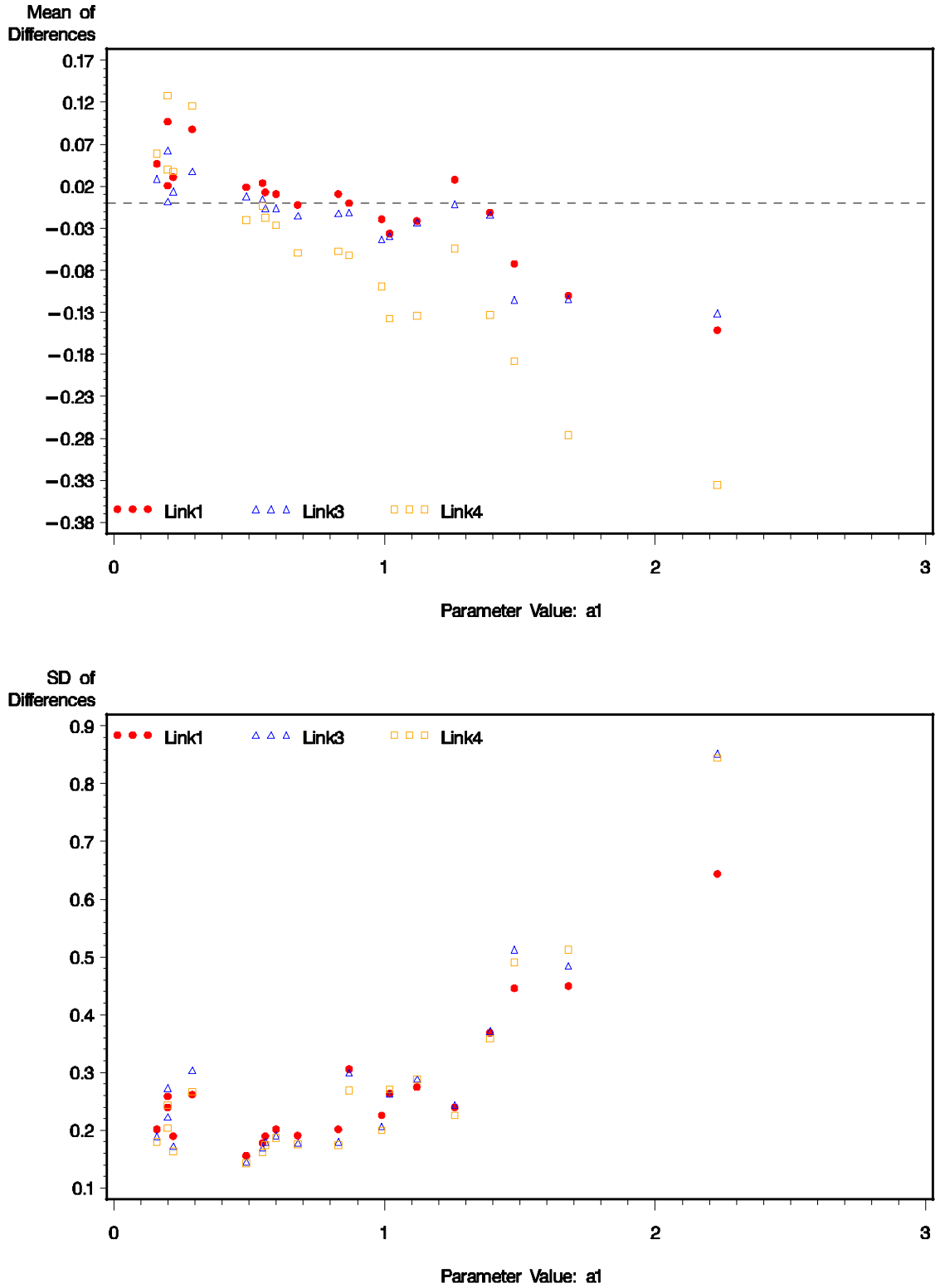
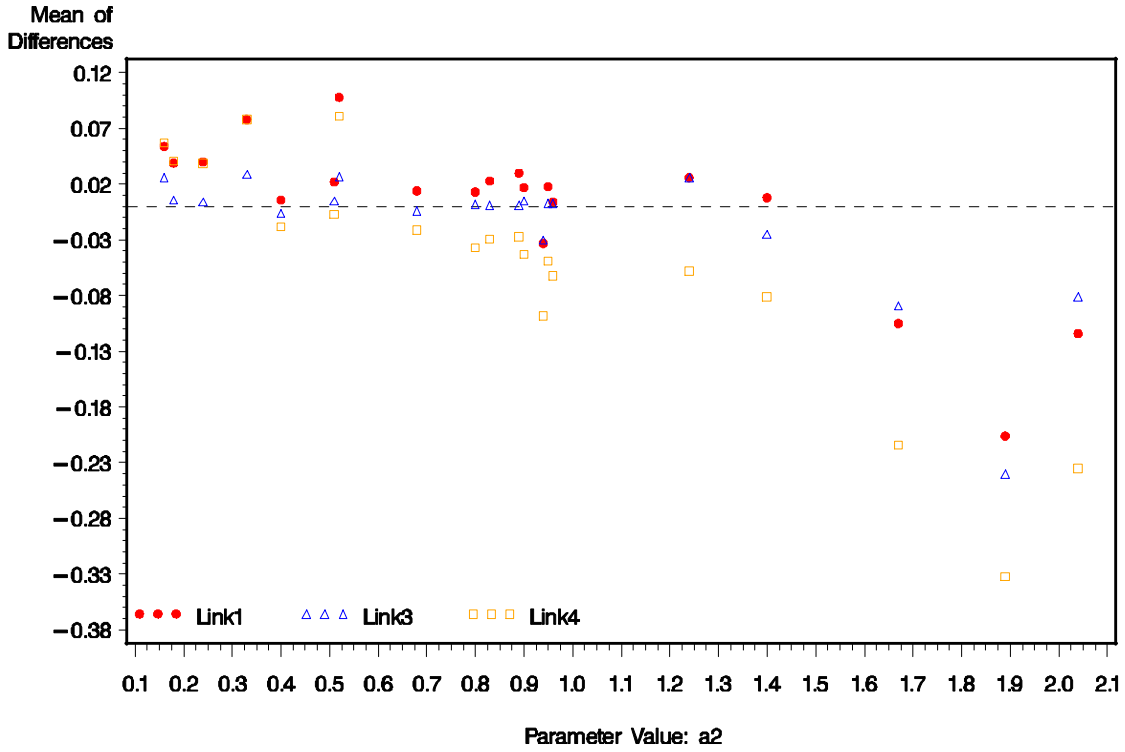


Figure 3-18. Linking accuracy and stability and item parameter values: COM, n=20, N=1000, G3



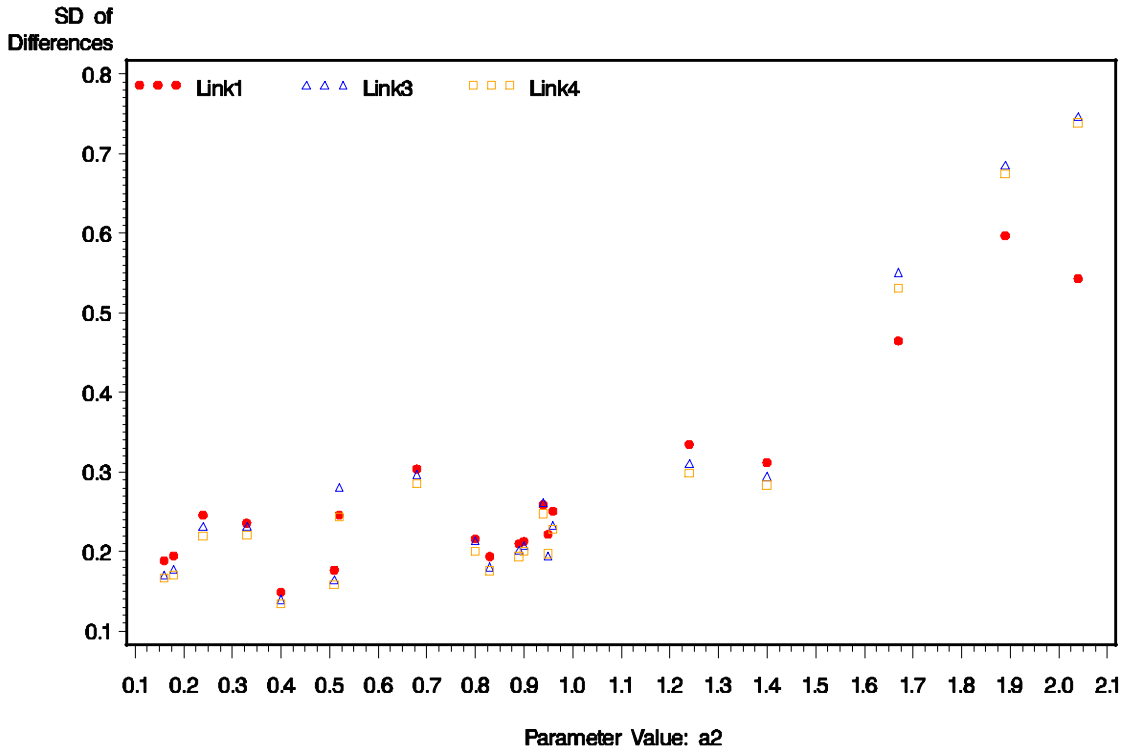
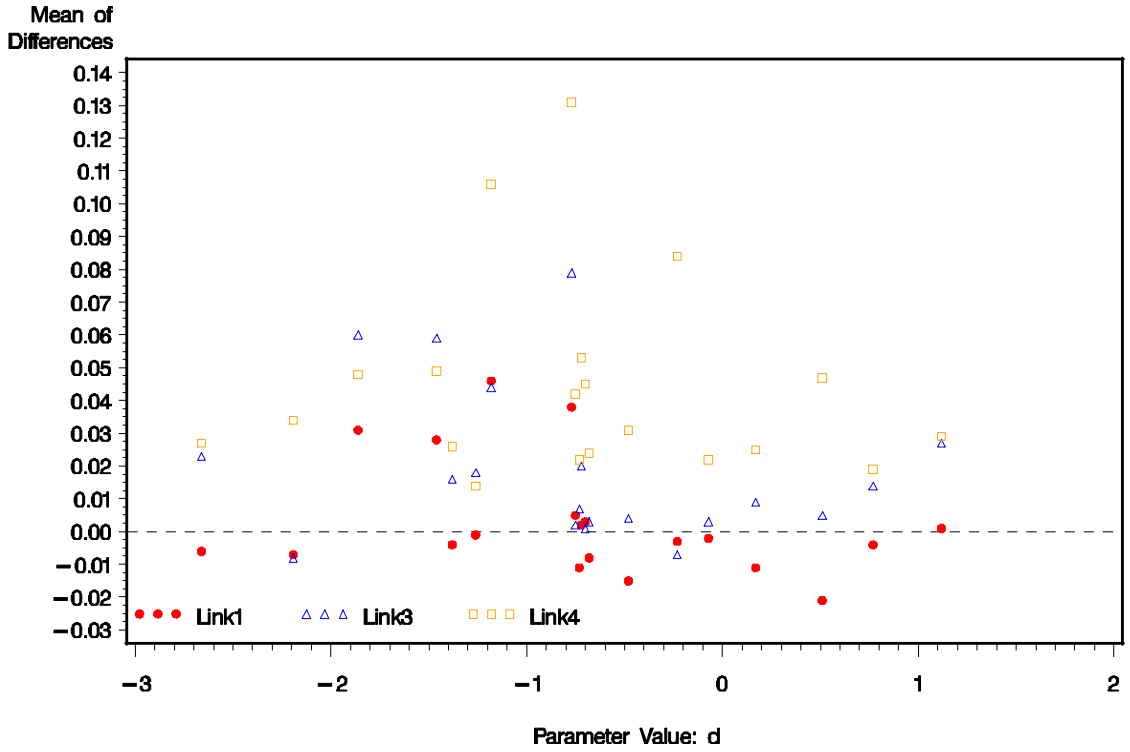


Figure 3-18. Continued



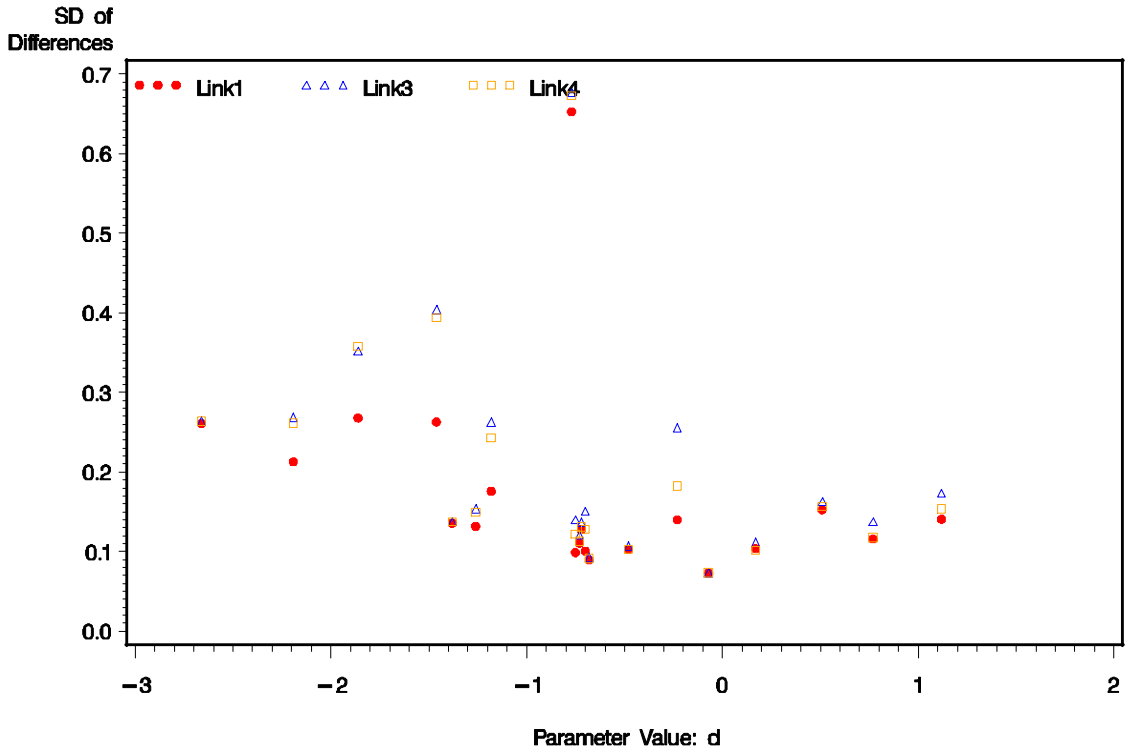


Figure 3-18. Continued

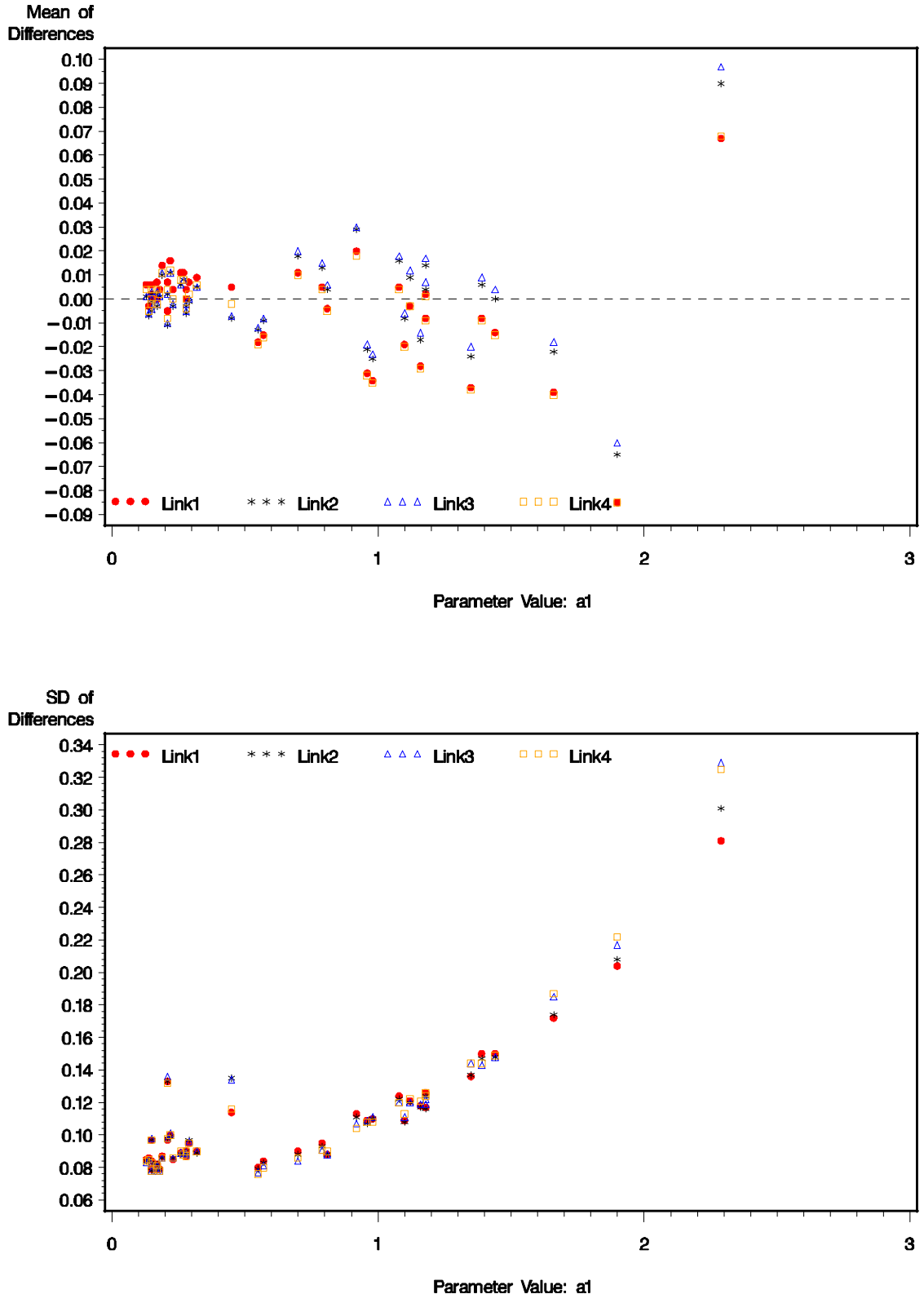
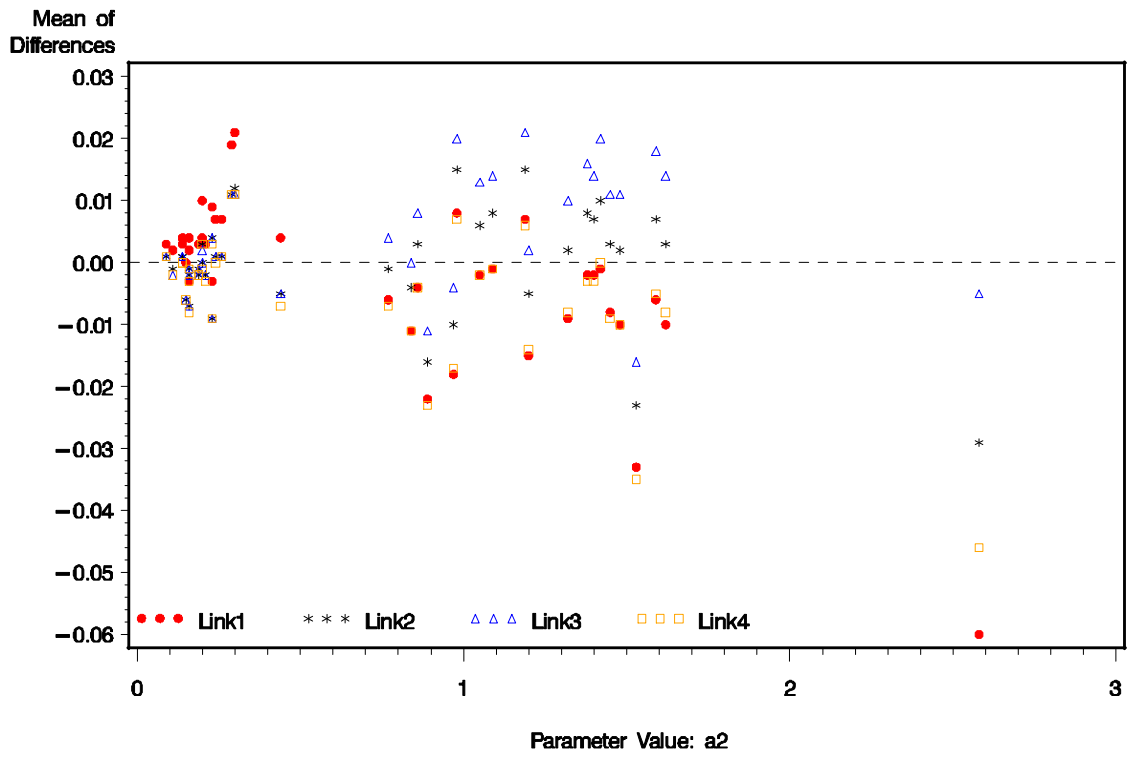


Figure 3-19. Linking accuracy and stability and item parameter values: APP, n=40, N=2000, G4



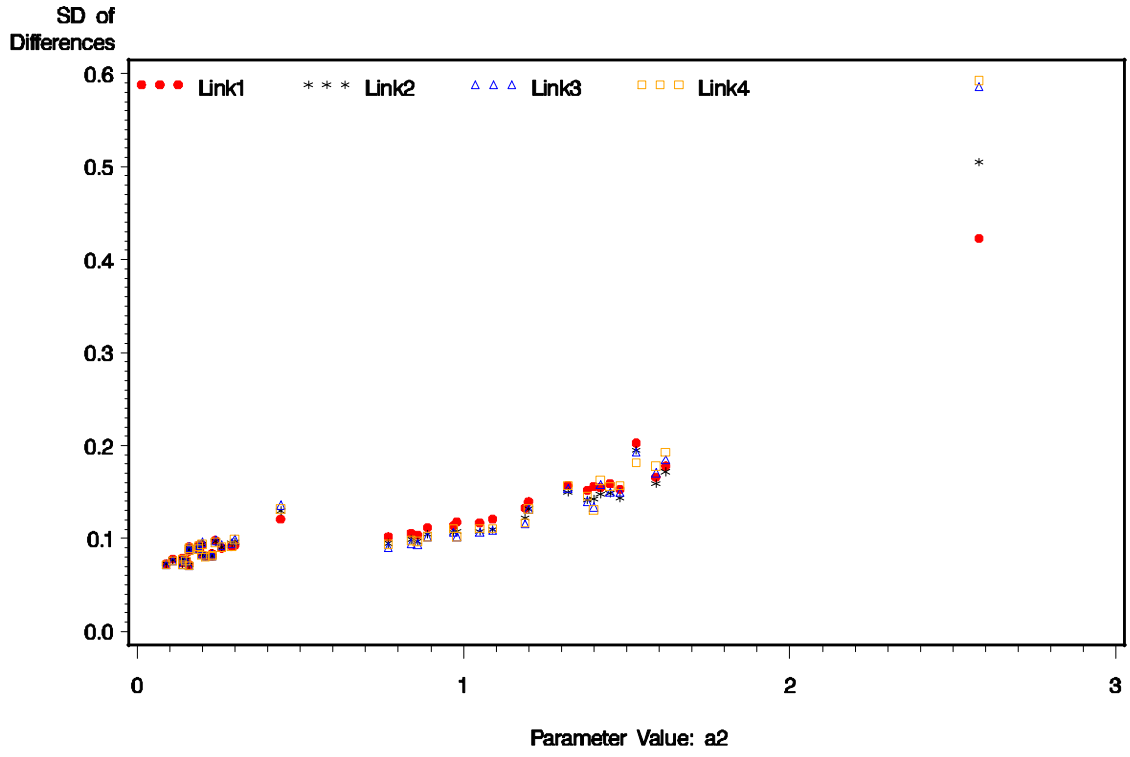
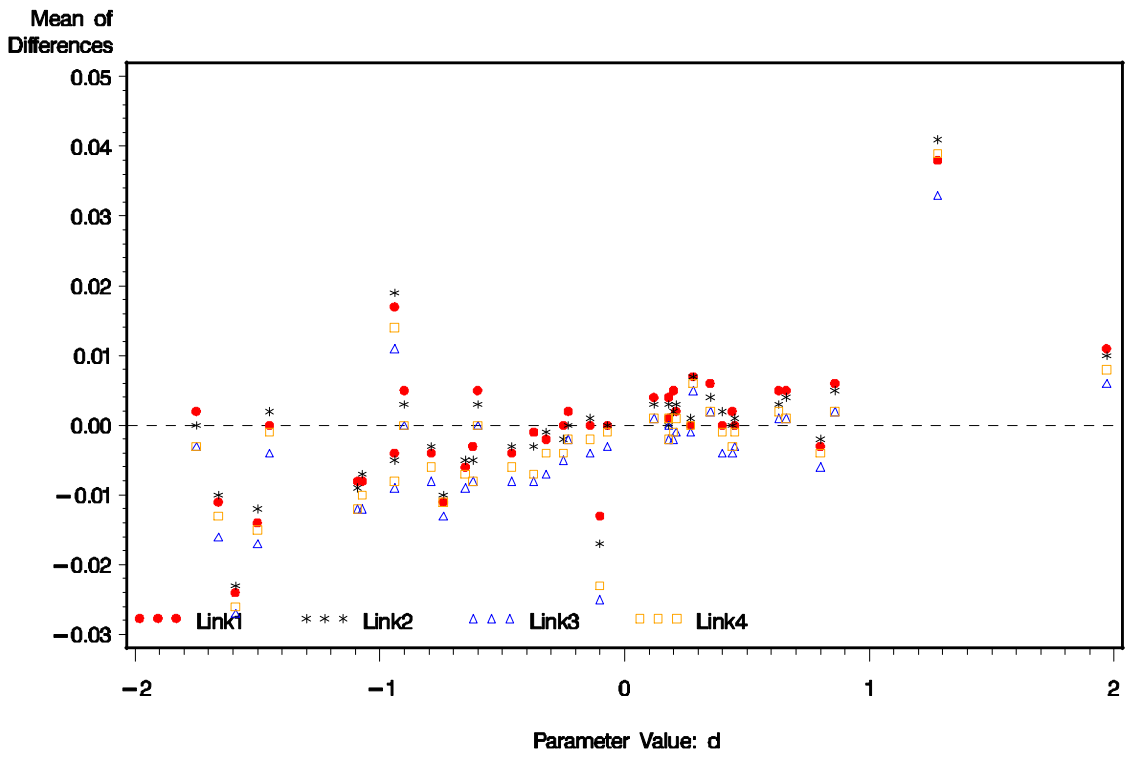


Figure 3-19. Continued



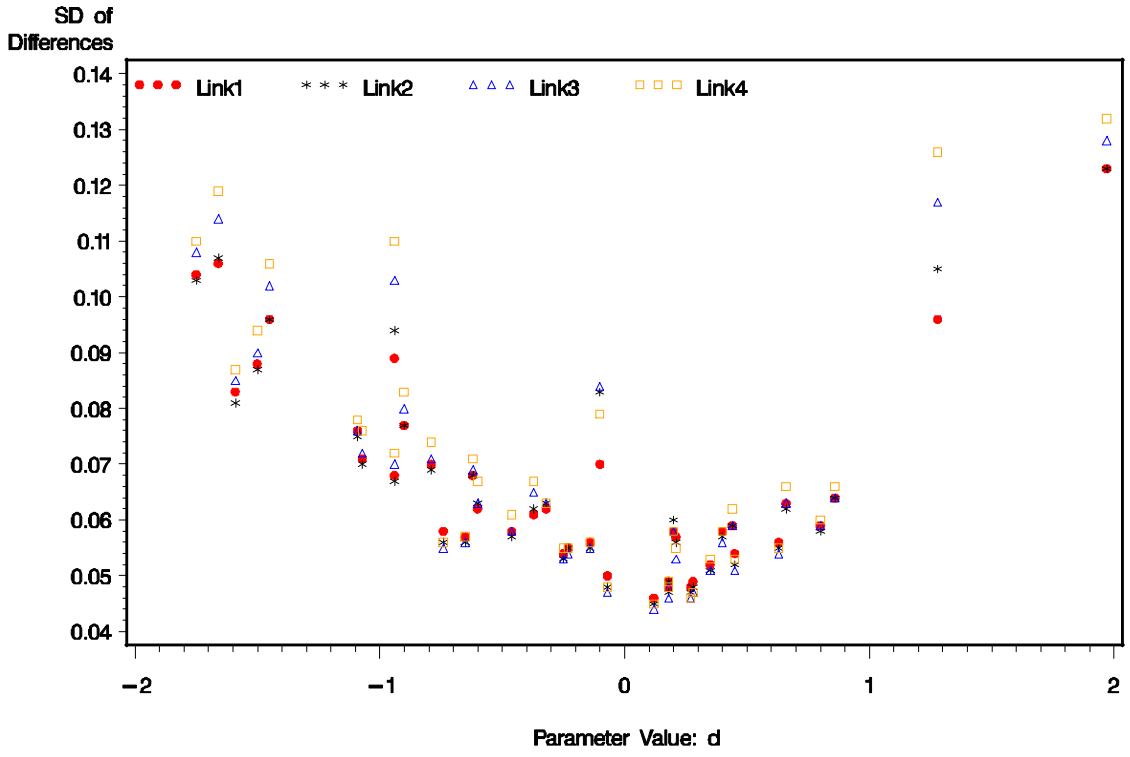


Figure 3-19. Continued

CHAPTER 4
DISCUSSION

By using simulated data, the performance of the four multidimensional IRT scale linking methods was evaluated under different testing conditions, which include different test structures, test lengths, sample sizes, and ability distributions. The results illustrated in Chapter 3 suggest that test structure had a strong influence on the performance of the four linking methods. For approximate simple test structure, each of the four linking methods worked approximately equally well under all testing conditions. For complex test structure, the equated function method did not work well under any testing conditions; the performance of other three linking methods depended on different testing conditions; the direct method was the best linking procedure for most testing conditions. In addition, the item parameter values influenced the linking performance. The results are discussed in this chapter by seven sections.

Results from Previous Studies

Theoretically, there are at least two main components in linking errors: error caused by parameter estimation and error produced by scale transformation (Li, 1997). A simulation study (Kaskowitz & Ayala, 2001) found that linking was more accurate when there was less error in the item parameter estimates. Therefore, it is important to review previous studies about IRT parameter estimation and linking accuracy under different testing conditions, although it is difficult, if not impossible, to decompose the parameter estimation error from the linking error in testing practice.

Based on Li's review (1997), the following factors can cause error in parameter estimation in IRT: (a) Examinees' ability distribution. Item difficulty for easy and hard items will not be well estimated when the examinees are normally distributed around their mean; Examinees with ability levels above or below the item difficulty are more informative for estimating item

discrimination parameter; (b) Item parameter value. Item difficulty parameters that are small or large and discrimination parameters that are small or large produce larger estimation error; (c) Sample size. Larger sample sizes reduce estimation error. However, the standard error of parameter estimates depends on the combined effect of these factors (Thissen & Wainer, 1982).

Using the bias and RMSE between the transformed linking parameter estimates and the true linking parameters across replications as the criterion, Li (1997) found that the linking accuracy of his three methods improved as sample size or test length increases. In the second study, Li (1997) used the bias and RMSE between the transformed item parameter estimates and the true parameter values across replications as the criterion and found that one of his linking methods (e.g., the combination of procrustean rotation approach for dimensional transformation, the ratio of trace procedure for dilation, and the least square procedure for translation) produced accurate linking of items. In addition, the positively skewed distribution of the second dimension in equated group did not negatively influence the linking accuracy and stability.

To evaluate the performance of the four multidimensional IRT scale linking methods, Oshima et al. (2000) used different criteria, including mean and standard deviation of the linking parameter estimates over 20 replications, bias and mean square error (MSE) between the estimated and true linking parameters, correlation and mean absolute difference of linking parameter estimates across different methods, and minimized function values by different methods. The results indicate that: (a) The direct method and equated function method tended to yield similar linking results and the test characteristic function method and item characteristic function method tended to produce similar results, (b) the test and item characteristic function methods were more accurate and stable than the other two methods, and (c) the accuracy and stability decreased as ability differences between the groups increased.

Min (2003) used the bias and RMSE between transformed item parameter estimates and the initial item parameters across the common items as the criterion to compare Li's (1997) composite procedure, Oshima and colleagues' test characteristic function method, and Min's extended composite procedure. Based on the repeated measures analysis of variance for bias and log transformed RMSE, The author found that: (a) The ability distribution, test structure, and linking method accounted for large portion of the variation in bias for discrimination parameter estimates but only linking method was an important factor for the variation in bias of difficulty parameter estimates, (b) the sample size, ability distribution, and linking method were important for linking stability of discrimination parameter estimates and sample size and linking method were critical for linking stability of difficulty parameter estimates, (c) as the sample size became larger and the two groups were more similar, the linking results became more accurate and stable, and (d) linking methods had significant interaction effects with testing conditions. In sum, the linking methods and the three testing conditions, e.g., the ability distribution, test structure, and sample size, significantly affected the linking accuracy and stability.

Effects of Different Test Structures

In the present study, the performance of all four linking methods worked much better for APP than for COM tests. This is consistent with the fact that the test structure and item parameters are typically more easily and accurately estimated for APP test than for COM test. As Tate (2003) found and discussed in a study comparing different estimation methods, including NOHARM, for assessing the test structure of item responses, the default rotation methods in exploratory analysis are usually developed to transform the initial solution to simple structure, therefore the procedures may not always successfully describe non-simple test structure. A study (Gosz & Walker, 2002) comparing the performance of TESTFACT and NOHARM found that the item parameter estimation of NOHARM depends heavily on the number of bi-dimensional

items in the test with its better performance for fewer bi-dimensional items and worse performance for more bi-dimensional items. In addition, NOHARM is good at estimating items with very low values on one discrimination parameter and high values on the other discrimination parameter. In this study, for APP tests, all items had higher values on one discrimination parameter and lower values on the other discrimination parameter; for COM tests, half of the items had approximately similar values on both discrimination parameter values. An investigation of item parameter estimation for the simulated data used in the present study indicated that the NOHARM program provided better item parameter estimation for APP tests than for COM tests. The superior estimation for the APP tests is likely the source of the superior linking results for the APP tests.

However, in a study (Min & Kim, 2003) comparing Li's composite procedure and Oshima and colleagues' test characteristic function method under different testing conditions, no apparent linking difference was found between APP and COM tests (see Figure 2-7, Min & Kim, 2003). One possible reason is that the item's loadings on the two dimensions in COM tests in this study were more similar than those in Min and Kim's study. Specifically the heavily cross-loaded items in their study had the direction of 50° - 65° and 25° - 40°, and the direction of heavily cross-loaded items in this study was selected from a normal distribution with mean of 45° and standard deviation of 10°. According to the finding by Gosz and Walker (2002), the item parameter estimates for COM tests in this study were less accurate, so that the linking results were more different between APP and COM tests. Another possible reason is related to the different criteria used to describe the linking performance. This study used the percentage of items with different means and standard deviations between the item parameter estimates for the base group and the transformed item parameter estimates for the equated group over 500

replications to evaluate linking results. Min and Kim's study (2003) used the bias and RMSE between true parameter values and transformed parameter estimates over both 20 items and 50 replications, which may have difficulty in identifying the differential influence of APP and COM tests on the linking results.

As shown in the Results chapter, the linking results (except for equated function method) were very similar for APP and COM tests when the sample size became large (N = 2000). This may be related to the possible improved item parameter estimation for larger sample size for both APP and COM tests. However, the attribution of different linking performance for the two types of tests to estimation error needs to be investigated by more controlled studies in the future.

Effects of Different Test Lengths

It was illustrated in the last chapter that the linking results for long tests were typically better than those for short tests, which is consistent with Li's finding (1997) that the linking accuracy of his three methods improved as test length increases. This result was not unexpected since more items can provide more information to set up the linkage between the scales for the base and equated groups. The positive effect of large number of items on linking and equating performance has already been found in various unidimensional equating conditions (Budescu, 1985; Fitzpatrick & Yen, 2001; Kaskowitz & Ayala, 2001; Kim & Cohen, 2002; Peterson, Cook, & Stocking, 1983; Swaminathan & Gifford, 1983; Wingersky, Cook, & Eignor, 1987). Therefore, this effect of the number of items can be extended from unidimensional to multidimensional linking and equating situations.

However, there was an exception that the linking results for short COM tests were better than those for long COM tests when the sample size was small (N = 500). Li could not find this exceptional result because he used sample sizes of 1000, 2000, and 4000 in his study. One possible reason for this exceptional result is that small sample size was not large enough to

produce accurate item parameter estimates for long test because more item parameters needed to be estimated, which accordingly affected the linking performance for long COM tests. Therefore, the strength of large number of items in scale linking and equating depends on the quality of the item parameter estimation, which in turn requires enough sample size. Lord (1980) stated that it is test length in combination with sample size that affects the quality of parameter estimates. Compared with unidimensional IRT models, a larger number of examinees are required for MIRT calibration because more parameters need to be estimated.

In addition, this study found that the effects of test length on scale linking performance also depended on the ability distributions for the two groups. As described in Results chapter, the long test (n = 40) did not improve linking performance when the means of ability distributions were different for base and equated groups for COM test when the sample size was 1000. This phenomenon confounded the general positive effect of large number of items on linking results. Klein and Kolen's study (1985) suggests that test length has little effect on the equating quality when groups are similar in ability, but becomes very important when two groups differ in ability level. They found that a larger number of common items did improve equating when groups were dissimilar. However, the exceptional result from this study mentioned above did not confirm their finding. Further studies are needed to examine the conflicting findings by controlling more conditions.

Effects of Different Sample Sizes

Based on the results from this study, the effects of sample size were very obvious and straightforward. Generally speaking, the linking accuracy and stability improved with the sample size increasing. This is consistent with the fact that large sample size can improve item parameter estimates. The same pattern was also found in the other two multidimensional scale linking studies (Li, 1997; Min & Kim, 2003). In addition, the positive effect of large sample size has

also been found in unidimensional linking and equating studies (Fitzpatrick & Yen, 2001; Hanson & Beguin, 2002; Kim & Cohen, 2002; Peterson, Kolen, & Hoover, 1989; Ree & Jensen, 1983).

However, the linking performance improved at different degrees for APP tests and COM tests. The performance of direct method, test characteristic function method, and item characteristic function method increased much more rapidly for COM tests than for APP tests when the sample sizes became larger. In fact, the linking results for APP tests were consistently good for different sample sizes. However, the linking results for COM tests were very different for different sample sizes, although the linking function improved with the sample sizes increasing. This result was not found in Min and Kim's study (2003). They showed similar effect of sample size on linking accuracy and stability for APP and COM tests (see Figure 2-7, Min & Kim, 2003). As we discussed for the effect of test structures on linking performance, this may be related to the different manipulations of COM test items and different evaluation criteria used in these two studies.

Effects of Different Ability Distributions

Based on this study, for all APP conditions and the COM conditions with a large sample size, between-group differences in ability distributions did not have a large influence on the performance of the four linking methods. For COM conditions with small and medium sample size (N=500, N=1000) between-group differences in mean ability had a negative influence on the linking results. It seems that mean difference was more important than variance difference. These results were consistent with what Oshima et al. found in their study using very similar ability conditions (see Table 5 and Figure 1, Oshima et al., 2000), although they did not divide tests into APP and COM tests. However, we need to be very cautious about the possible differential effect of mean and variance differences on scale linking in both studies because they

were controlled at different degrees, with mean difference at 0.5 and the variance difference at 0.2.

Based on the study by Min and Kim (see Figure 2-7, Min & Kim, 2003), it seems that the influence of ability distributions on scale linking by the test characteristic function method was approximately similar for APP and COM conditions (see the above explanation for possible reasons for this conflicting findings between their study and this study). However, they did find that the influence of between-group differences in ability distribution on scale linking depended on sample size, with less influence for large sample size (N = 2000) and more influence for small sample size (N = 500). This is consistent with the results from this study.

Li (1997) used a different manipulation of the between-group difference in ability distribution than was used in the present study: for the base group both ability distributions were normal; for the equated group one ability distribution was normal and the other was positively skewed. No negative effect was found on the linking performance by using his three methods. The reason may be that although the second ability had positively skewed distribution, the mean and standard deviation were still controlled at 0 and 1, which were the same as for the based group for the second dimension. It seems that mean and standard deviation were more important than the normality of the distribution. However, this conclusion needs to be confirmed for the MIRT linking methods.

Based on the research on unidimensional scale linking and test equating (see the review by Kolen and Brennan, 2004), the similarity between two groups of examinees affects linking and equating performance: the more similar the groups are, the more adequate the linking and equating will be; large differences between groups may produce significant problems. Based on results from multidimensional scale linking, this conclusion can be extended to the

multidimensional cases, but with cautious consideration of the interaction between ability distribution, test structure, and sample size.

Effects of Different Item Parameter Values

As mentioned in the first section, estimation of the item difficulty parameter is less accurate when the parameters are small or large, estimation of discrimination parameter is less accurate when discrimination parameters are small or large, and error in item parameter estimation affects scale linking performance. Therefore, linking quality is likely to be influenced by the item parameter values, especially by the extreme parameter values. This conceptual inference and conclusion were confirmed in this study: under most of the testing conditions, the linking results tended to be less accurate when the absolute item parameters had extreme values and less stable when the absolute item parameter values became large. This pattern of results was more apparent when (a) the test had approximate simple structure, (b) the sample size was larger, and (c) the linking performance for discrimination was evaluated.

The only other multidimensional scale linking study evaluating the effects of different item parameter values was conducted by Li (1997). Based on that study (see Figure IV-1-16, Li, 1997), the linking results for difficulty were more accurate and stable when the absolute item parameter values became larger; the linking results for discrimination did not change consistently with the item parameter values.

Therefore, the effects of different item parameter values on scale linking were more apparent for discrimination in this study and more obvious for difficulty in Li's study. This is reasonable given Min and Kim's (2003) conclusion that Li's method worked better than Oshima and colleagues' test characteristic function method (2000) for difficulty parameters and Oshima and colleagues' method worked better for the two discrimination parameters.

Performance of Different Linking Methods

The effects of test structure, test length, sample size, ability distribution, and item parameter values on scale linking performance were separately discussed above. However, these factors interacted with each other and had both main and combined effect on the performance of the four linking methods.

As summarized at the beginning of this chapter, generally speaking, all four linking methods worked approximately equally well under all testing conditions for approximate simple tests. For complex tests, the direct method was the best linking procedure; the item characteristic function method and the test characteristic function method were the second and third; the equated function method did not work well for complex tests. These results were based on the differences between the item parameter estimates for base group and the transformed item parameter estimates for equated group for the common items.

It is not entirely surprising that the direct method, which minimizes the sum of squared differences between the two sets of item parameter estimates over items, was the best one across different testing conditions because the evaluation criterion was consistent with the method. However, the equated function method estimates the linking parameters by minimizing the sum of squared difference between the means of the two sets of selected item parameter estimates in the test. It uses the accumulative information of some items. Therefore, it is possible that even though the mean parameter estimates were similar for the two groups, individual parameter estimates were not. In the same way, item characteristic function method uses the combined information of discrimination, difficulty, and ability item by item. The test characteristic function method uses the accumulative information of discrimination, difficulty, and ability over all items in the test. Therefore, item characteristic function method was better than test characteristic function method using the criterion based on difference between item parameter estimates.

Why do the four linking methods worked equally well for approximate simple tests but differentially poor for complex test? One of possible reason is that there is complicated interaction between item parameter estimation error and the characteristics of the four linking methods. More simulation studies need to be conducted to differentiate the two types of effect on the performance of scale linking.

CHAPTER 5
CONCLUSIONS

The purpose of this study was to use simulated data to examine the performance of four multidimensional linking methods under different testing conditions. There were one hundred and ninety-two experimental conditions in this study: four linking methods (direct method, equated function method, test characteristic function method, and item characteristic function method) by two test structures (approximate simple test structure and complex test structure) by two test lengths (20 items and 40 items) by three sample sizes (500, 1000, and 2000), and by four different ability distributions between two groups (no difference, only mean difference, only variance difference, and both mean and variance difference). Five hundred replications were conducted for each of the experimental conditions. The linking performance evaluation was based on the differences between the item parameter estimates for base group and the transformed item parameter estimates for equated group for the common items. The mean and standard deviation of the differences across the 500 replications were computed to examine the accuracy and stability of the four linking methods.

Conclusions

- **Conclusion 1:** The performance of the four linking methods. Generally speaking, the direct method was the best linking method; the item characteristic function method and test characteristic function method were the second and third best method; the equated function method was the last method. However, their linking performance depended on the following testing conditions.

- **Conclusion 2:** The effects of test structure. For approximate simple test structure, each of the four linking methods worked approximately equally well for all testing conditions; For complex test structure, the equated function method worked poorly under all testing conditions; the performance of the other three linking methods depended on other testing conditions; the direction method was the best method for most testing conditions.

- **Conclusion 3:** The effects of test length. The linking performance for long tests was typically better than that for short tests except for complex tests when the sample size was small.

- **Conclusion 4:** The effects of sample size. The linking performance improved when the sample size became larger, especially for complex tests.

- **Conclusion 5:** The effects of ability distribution. Quality of linking performance declined when there was difference in ability distribution between the two groups, especially for complex tests; however, it seems that a between-group difference in the means was more important than a difference in the variance.

- **Conclusion 6:** The effects of item parameter values. Under most of the testing conditions, the linking results for the discrimination parameter tended to be less accurate and less stable when the item parameter had extreme values. The linking accuracy for the difficulty parameter was not dependent on the item parameter values. The linking stability for the difficulty parameter depended on the item parameter values only when the sample size was large. Then, the linking results were less stable when the item parameter had extreme values.

Future Research

In this study, there are a number of limitations, which should be considered for making the conclusions described above. For example: (a) Although the item parameters for short and long approximate simple tests and complex tests were randomly created in the same way and from the same distributions, they did not have the same exact values. This should be considered when comparing the linking results for the four types of tests; (b) The test structure was not constructed by randomly arranging the items in the test. For the approximate simple test, the first half of the items had higher discrimination values for the first ability and lower values for the second ability, and the second half of the items had lower discrimination values for the first ability and higher values for the second ability. The equated function method in this study used the means of the first half of items (all with lower or higher values), second half of items (all with lower or higher values) as the function to estimate the linking parameters. This may affect the linking performance of the equated function method; (c) This study used the differences between the item parameter estimates for base group and the transformed item parameter estimates for equated group as the criterion, which is consistent with the minimized function of

the direct method and accordingly may favor this method. The linking performance should be evaluated using other criteria which are consistent with the other methods to examine the possible dependence of the results on the criteria used. These criteria include the differences between the means of the selected item parameter estimates obtained from the two groups, the differences between the test characteristic functions for a given range of ability, and the differences between item characteristic functions for a given range of ability.

As mentioned in the first chapter, the development of multidimensional linking methods is just at the infancy stage and more research is needed to obtain definitive results. Therefore, a substantial research needs to be conducted to explore and evaluate different procedures for multidimensional scale linking. Here are some future research topics on multidimensional IRT scale linking.

First of all, different specific procedures within each of the four linking methods need to be explored, compared, and evaluated so that the best method can be chosen for some specific purpose. For example: (a) For the test characteristic function and item characteristic function methods, how should the theta region or levels be chosen? Should we use the equally spaced grid theta method or empirical theta method? If we choose empirical theta method, which examinee group, base group, equated group, or combined group, should be used? Which method is better? Should we give different weights to different theta regions and how to choose different weights? (b) For equated function method, which item parameter estimates should be used? What characteristics should be considered to choose the appropriate sets of items? What function should be used to produce good linking performance?

Second, what kind of criteria should be used to evaluate the performance of different linking methods? Within the multidimensional IRT linking and equating studies, different

criteria have been used. Even within one study, different criteria have been used. For example, Li (1997) used bias and RMSE between the transformed linking parameter estimates and the true linking parameters across replications in his first study and then used bias and RMSE between true item parameter values and the transformed item parameter estimates and ability recovery in his second study. Oshima et al. (2000) used mean and standard deviation of the linking parameter estimates over 20 replications, bias and MSE between the estimated and true linking parameters, correlation and mean absolute difference of linking parameter estimates across different methods, and minimized function values by different methods. Min (2003) used bias and RMSE between transformed item parameter estimates and the initial item parameters across common items for the simulated data, and used the differences between the item parameter estimates for base group and the transformed item parameter estimates for equated group across the common items for the real data. Given these criteria, which one should we use for which purpose for scale linking? This is a critical issue in evaluating different methods.

Third, as we discussed in last chapter, there are at least two main components in linking errors: error caused by parameter estimation and error produced by scale transformation. The problem is how to differentiate the estimation error from the linking error when scale linking is conducted? To answer this question, many studies need to be conducted to evaluate the performance of different estimation programs for multidimensional IRT. In addition, some methods need to be developed to differentiate the estimation error from linking error and evaluate the effects of estimation error on the performance of scale linking.

Finally, the two approaches, multidimensional IRT approach and factor analysis approach, have different strengths and weaknesses in linking different scales. As Min and Kim (2003) found in their study that Li's method worked better for difficulty parameters and Oshima and

colleagues' method (e.g., test characteristic function method) worked better for the two discrimination parameters. Therefore, how to use the strengths of the two approaches to develop a combined method for multidimensional scale linking is an important topic in the future research.

APPENDIX
ACCURACY AND STABILITY FOR DIFFERENT LINKING METHODS

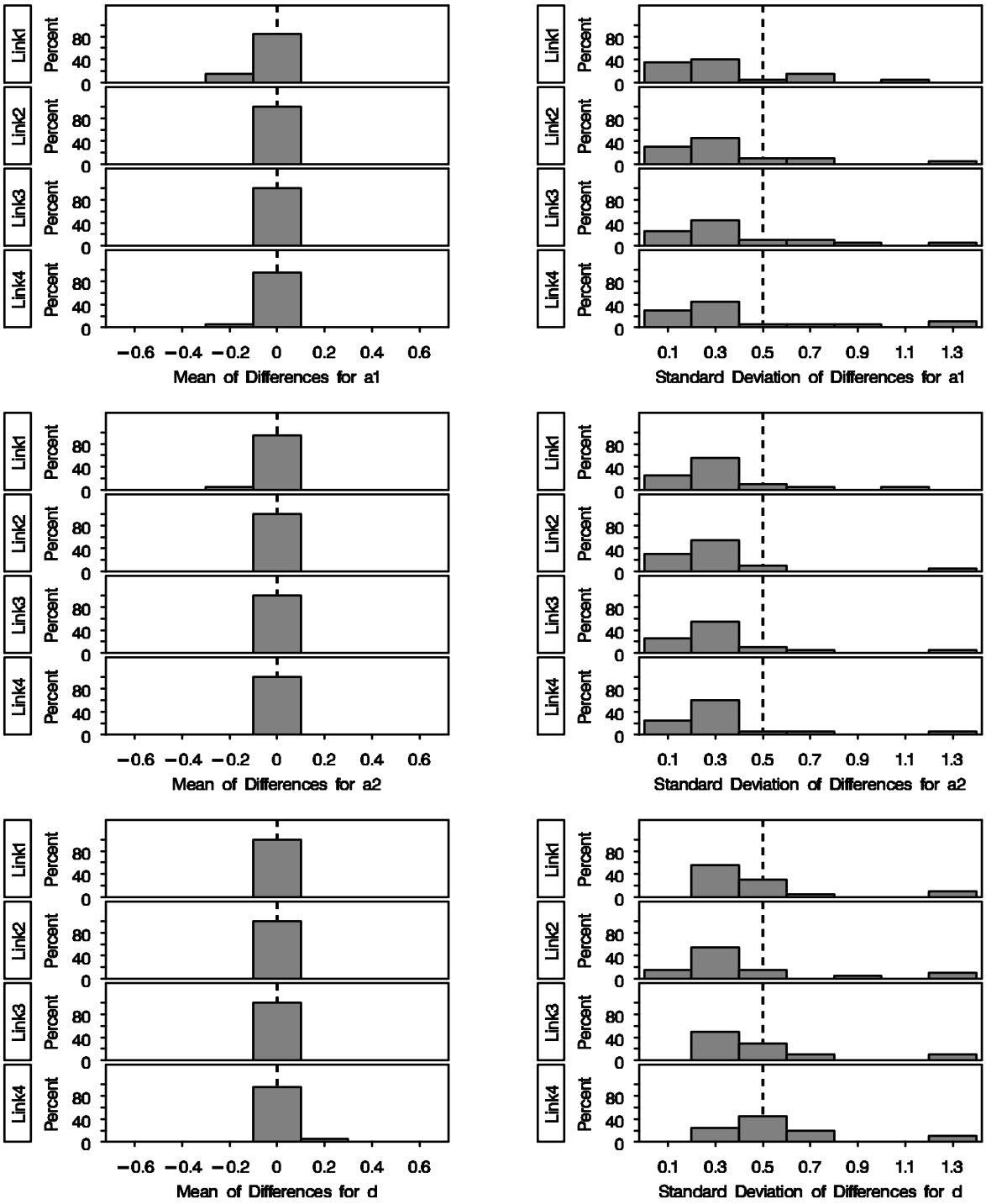


Figure A-1. Accuracy and stability for different linking methods (APP, n=20, N=500, G1)

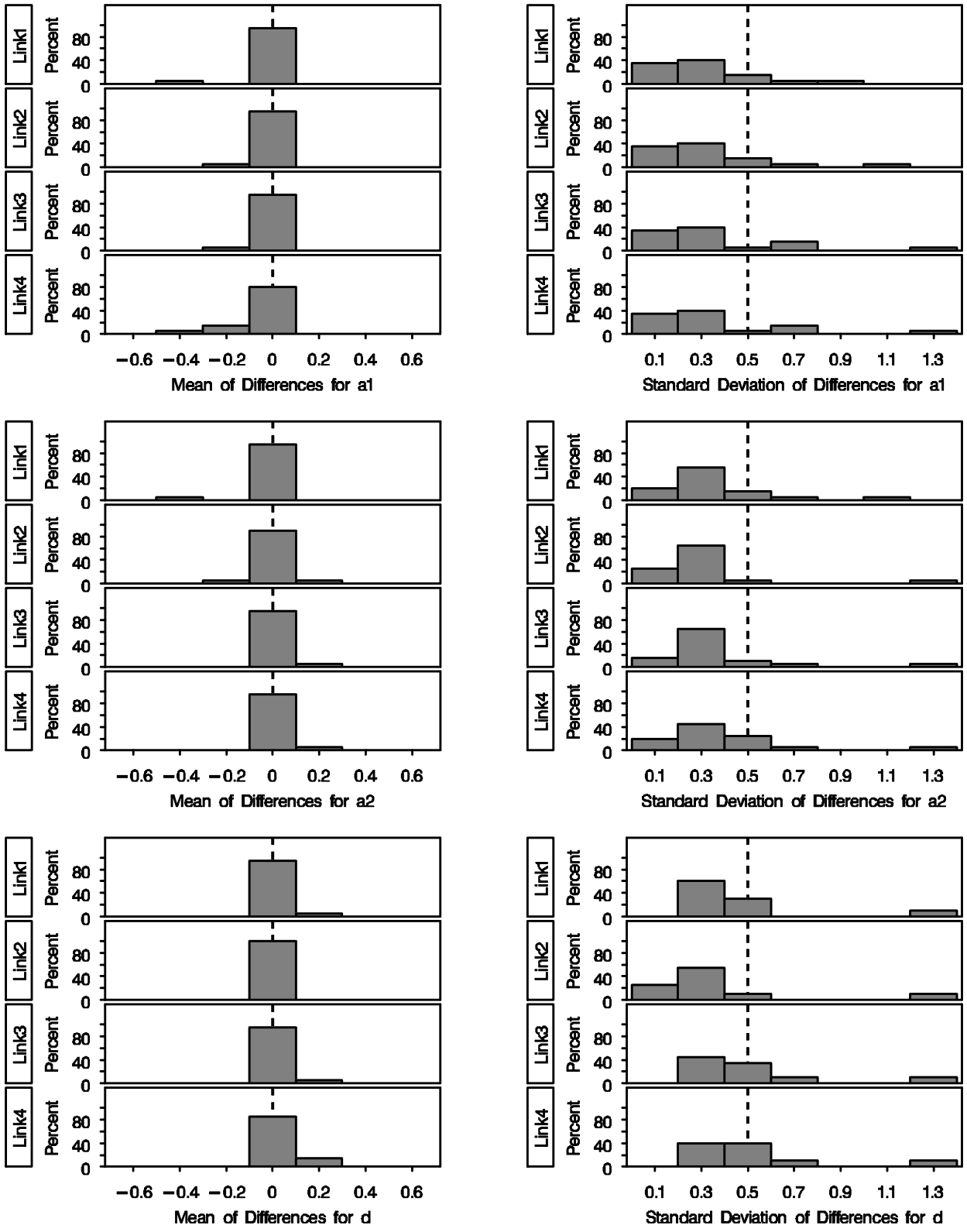


Figure A-2. Accuracy and stability for different linking methods (APP, n=20, N=500, G2)

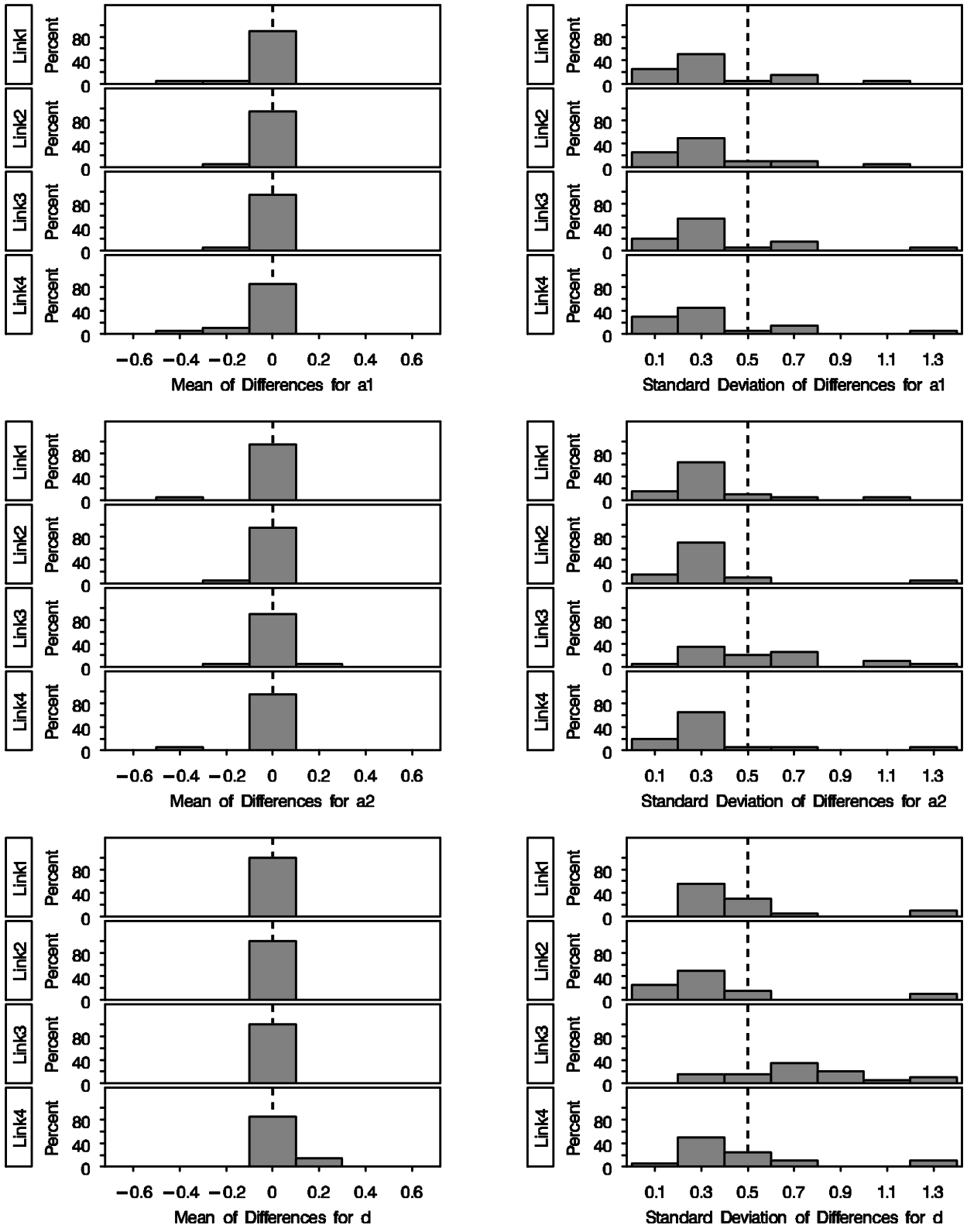


Figure A-3. Accuracy and stability for different linking methods (APP, n=20, N=500, G3)

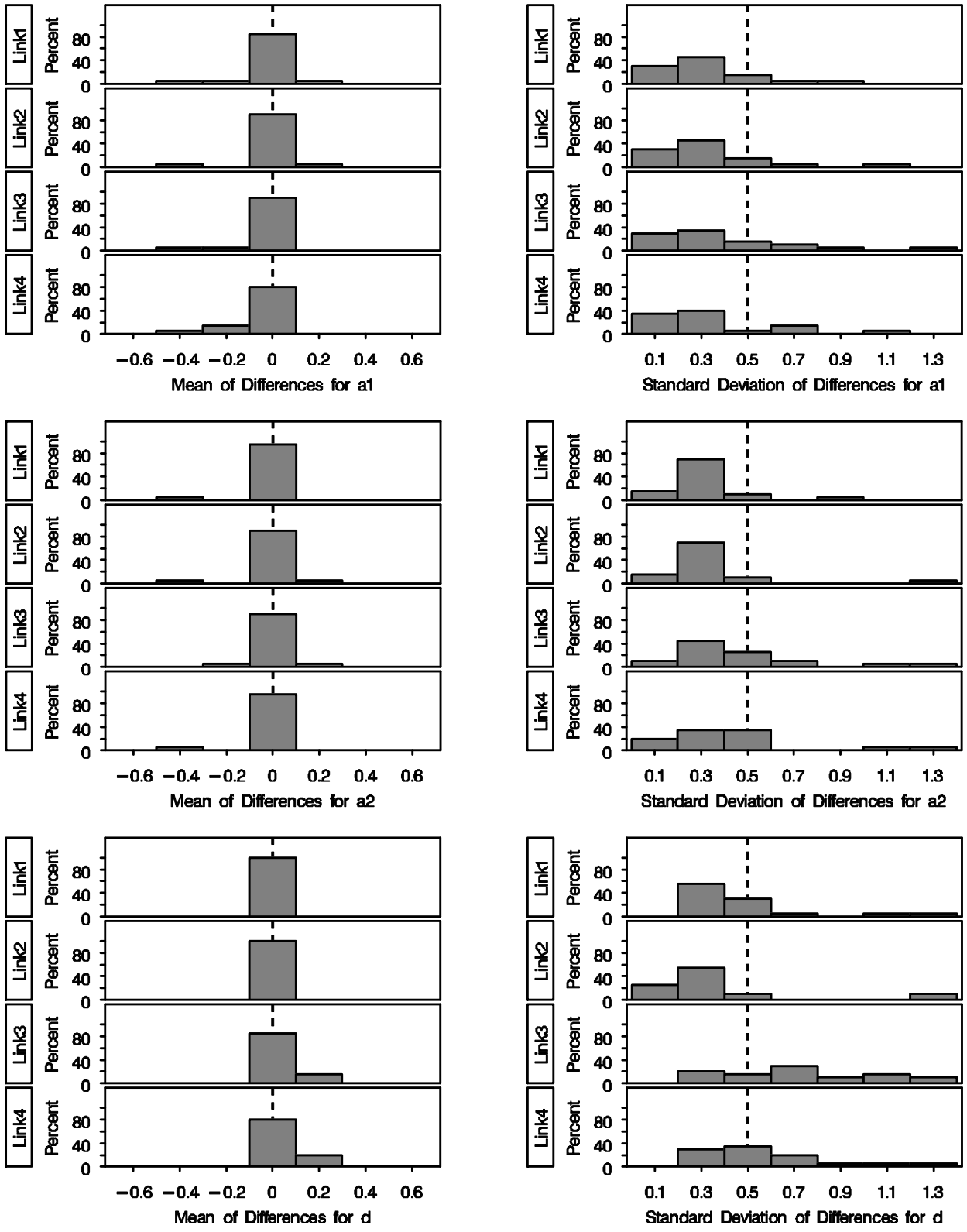


Figure A-4. Accuracy and stability for different linking methods (APP, n=20, N=500, G4)

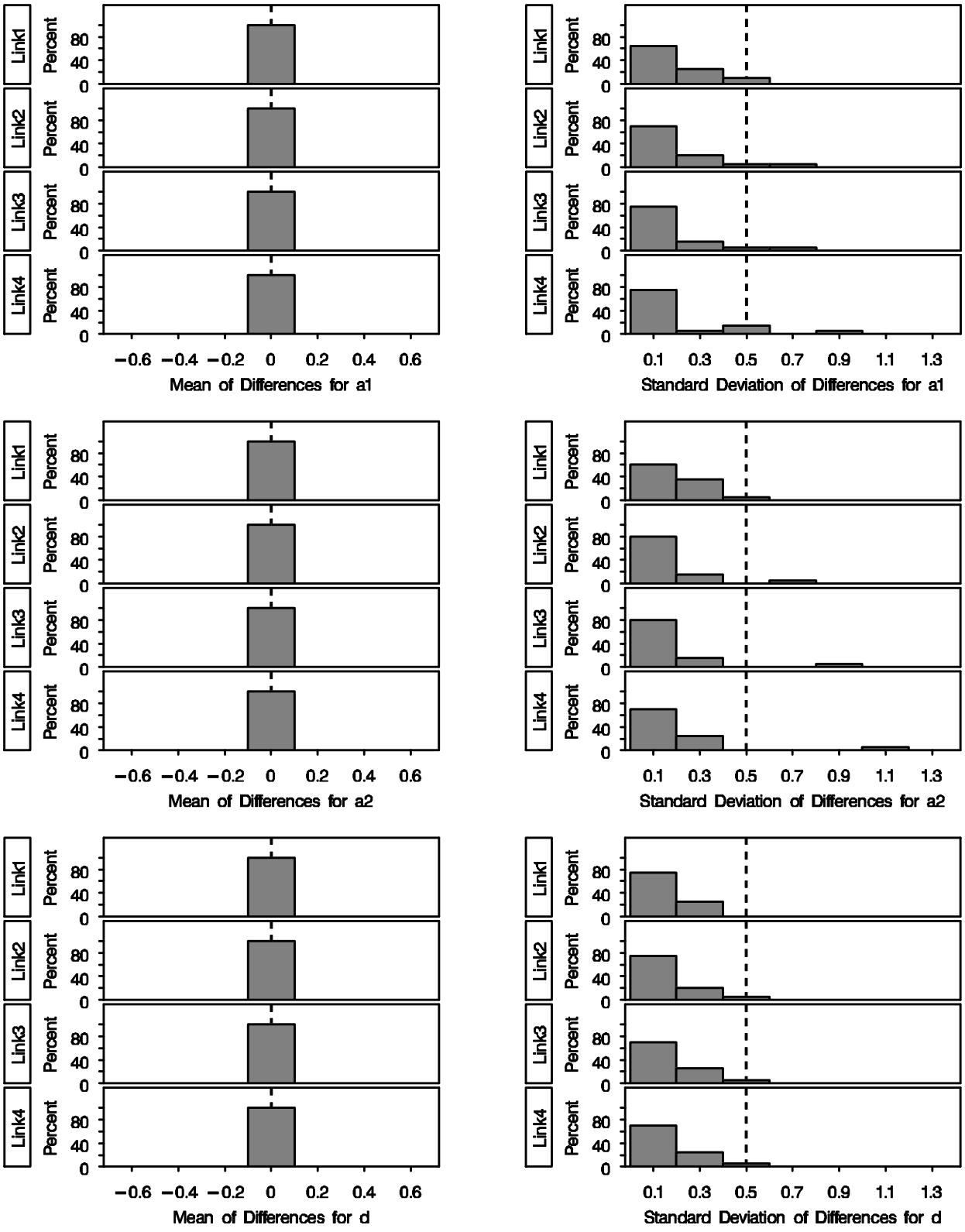


Figure A-5. Accuracy and stability for different linking methods (APP, n=20, N=1000, G1)

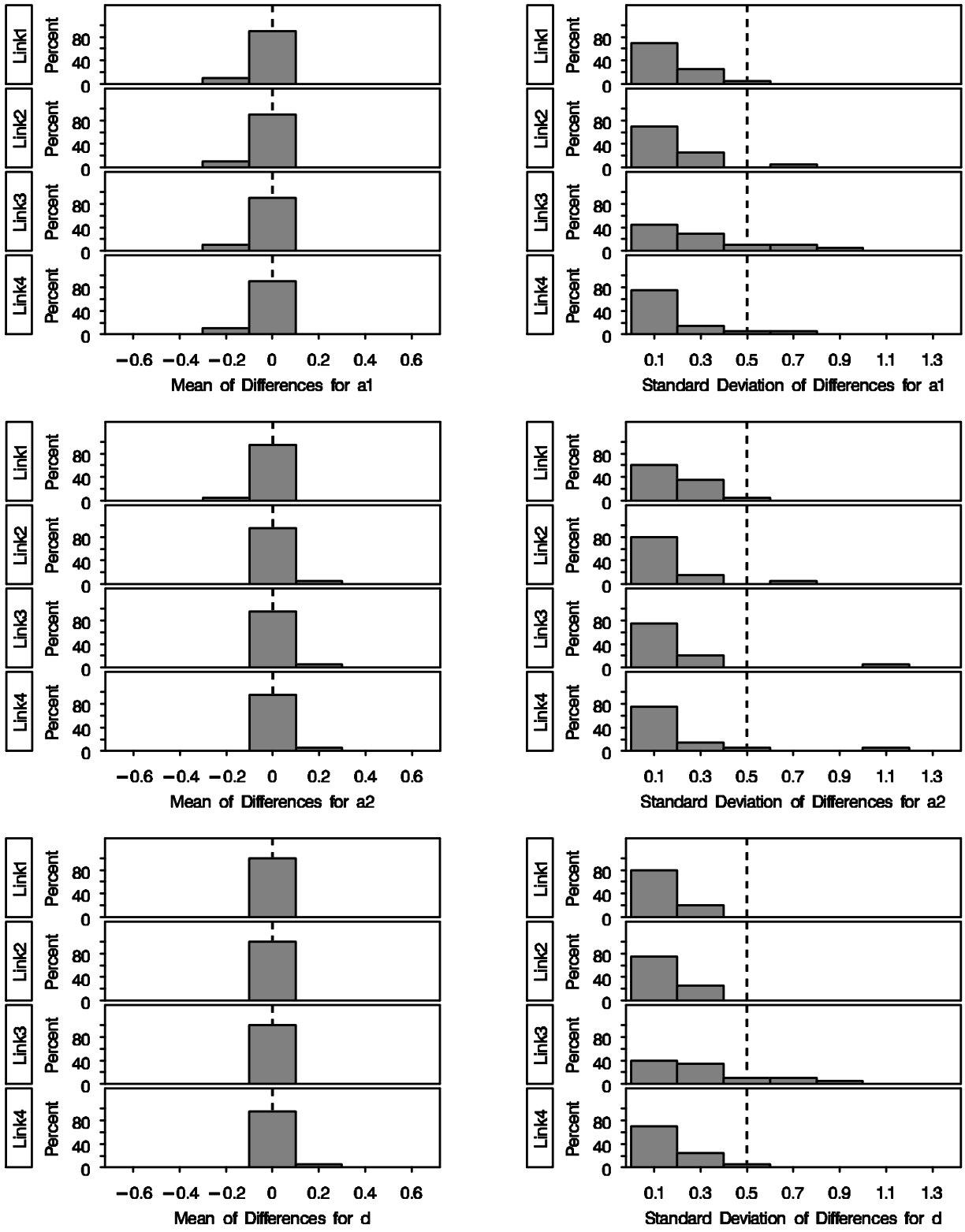


Figure A-6. Accuracy and stability for different linking methods (APP, n=20, N=1000, G2)

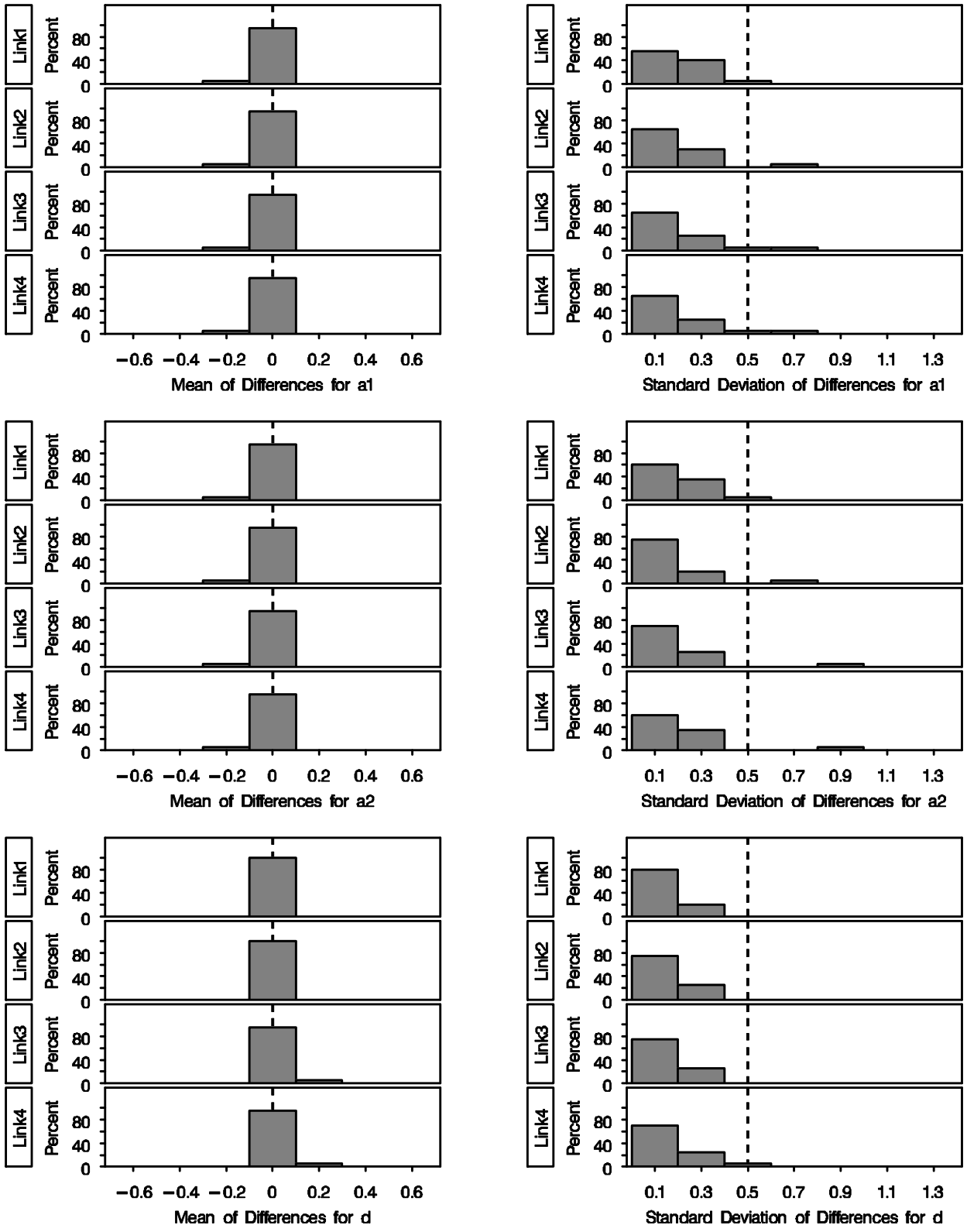


Figure A-7. Accuracy and stability for different linking methods (APP, n=20, N=1000, G3)

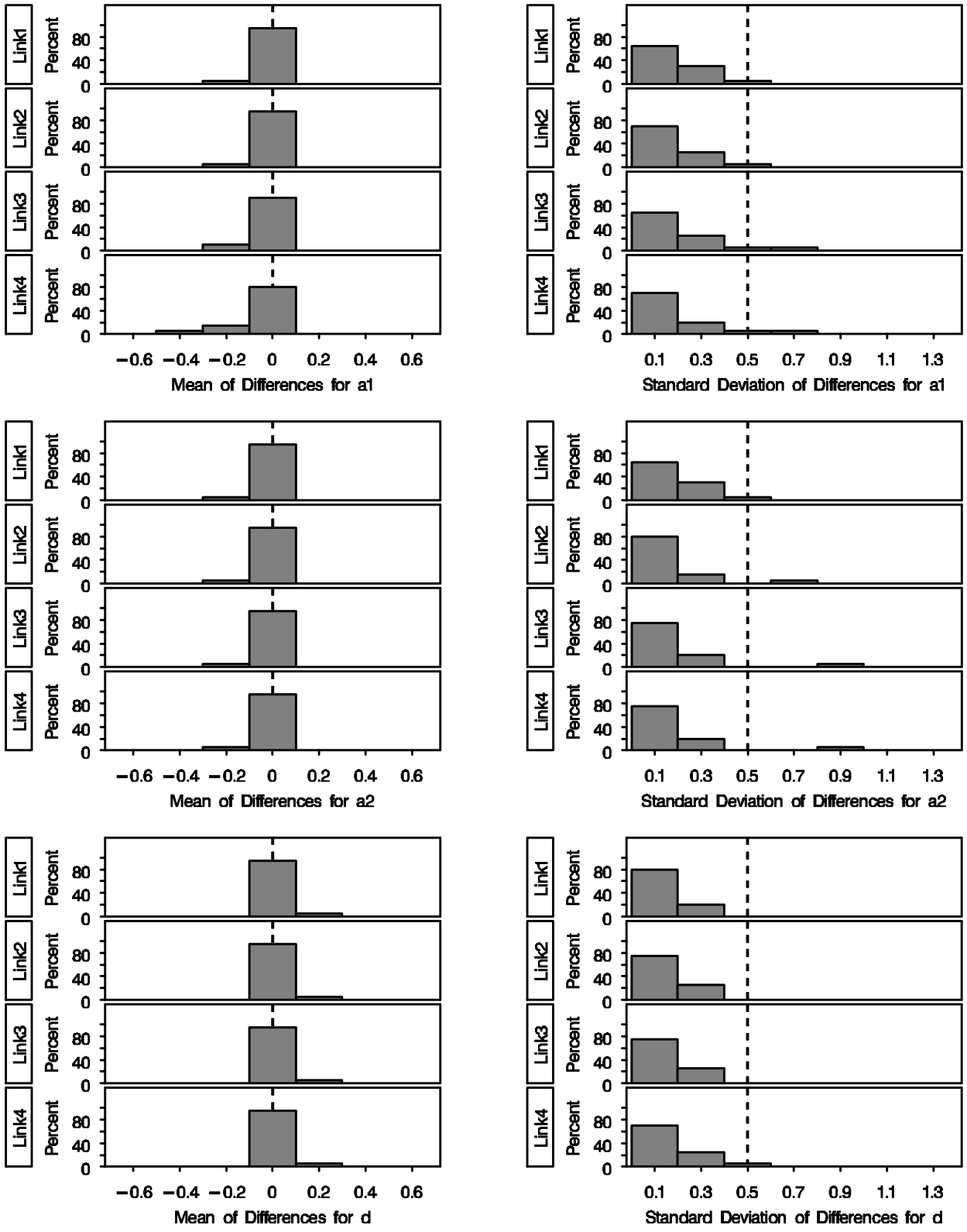


Figure A-8. Accuracy and stability for different linking methods (APP, n=20, N=1000, G4)

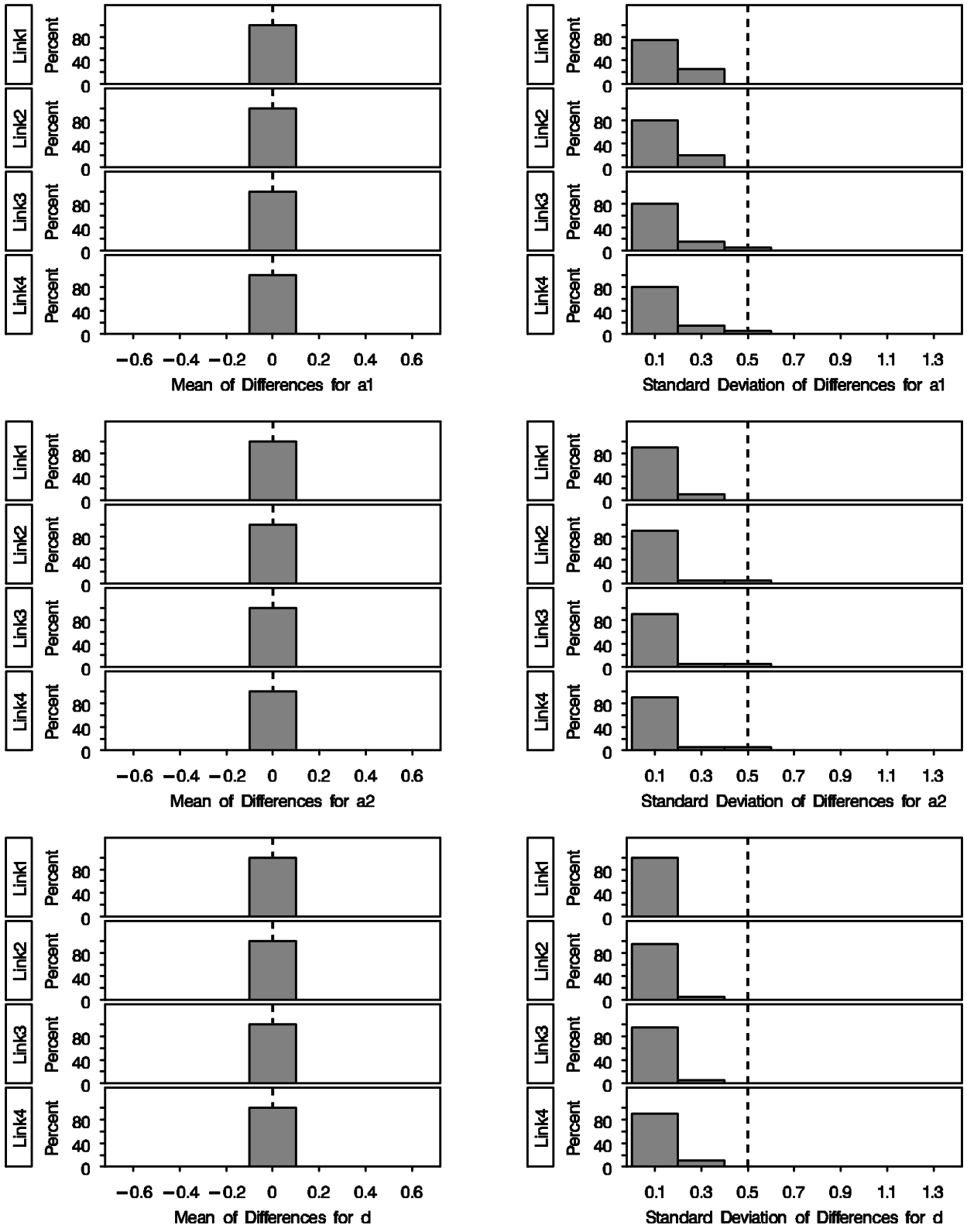


Figure A-9. Accuracy and stability for different linking methods (APP, n=20, N=2000, G1)

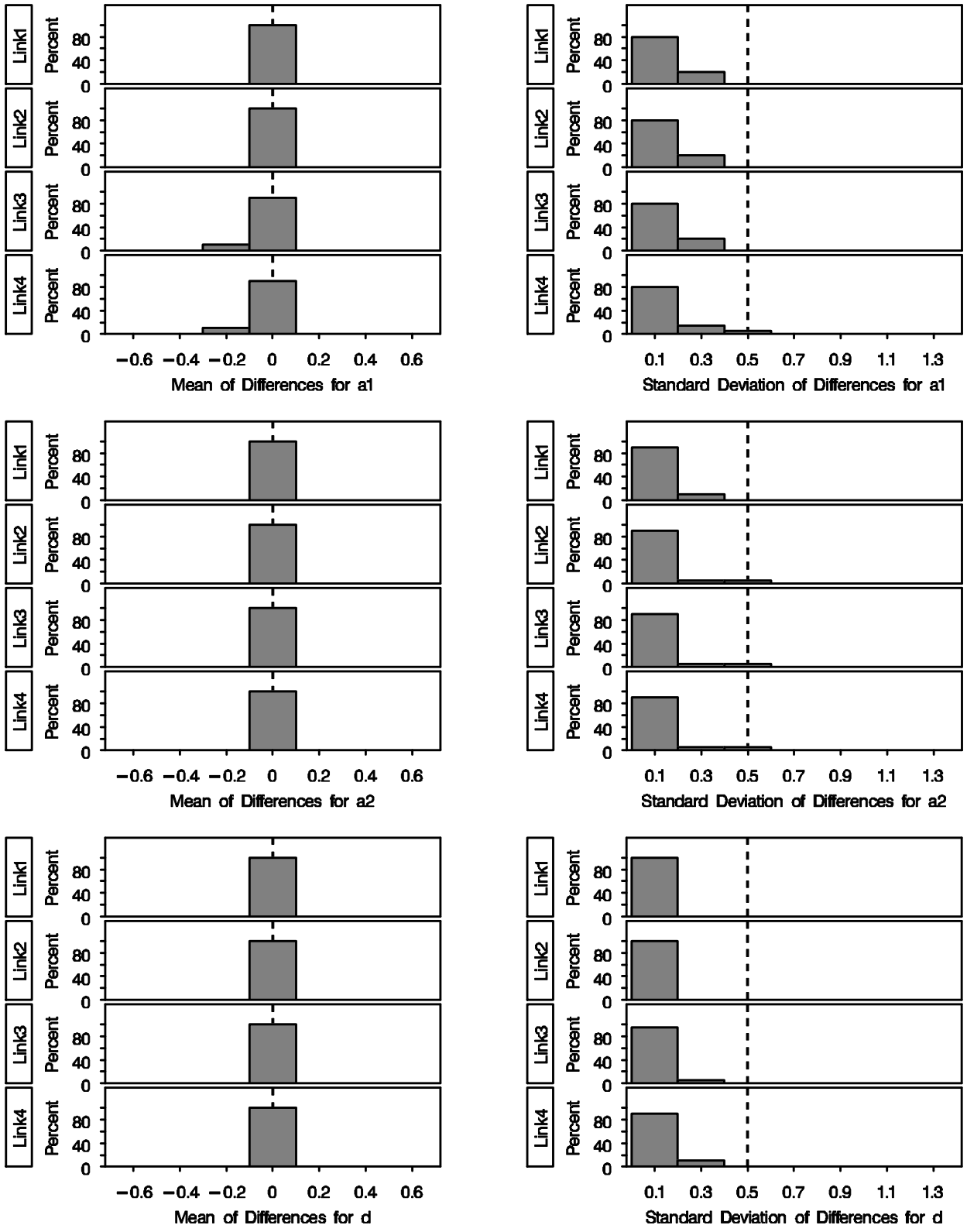


Figure A-10. Accuracy and stability for different linking methods (APP, n=20, N=2000, G2)

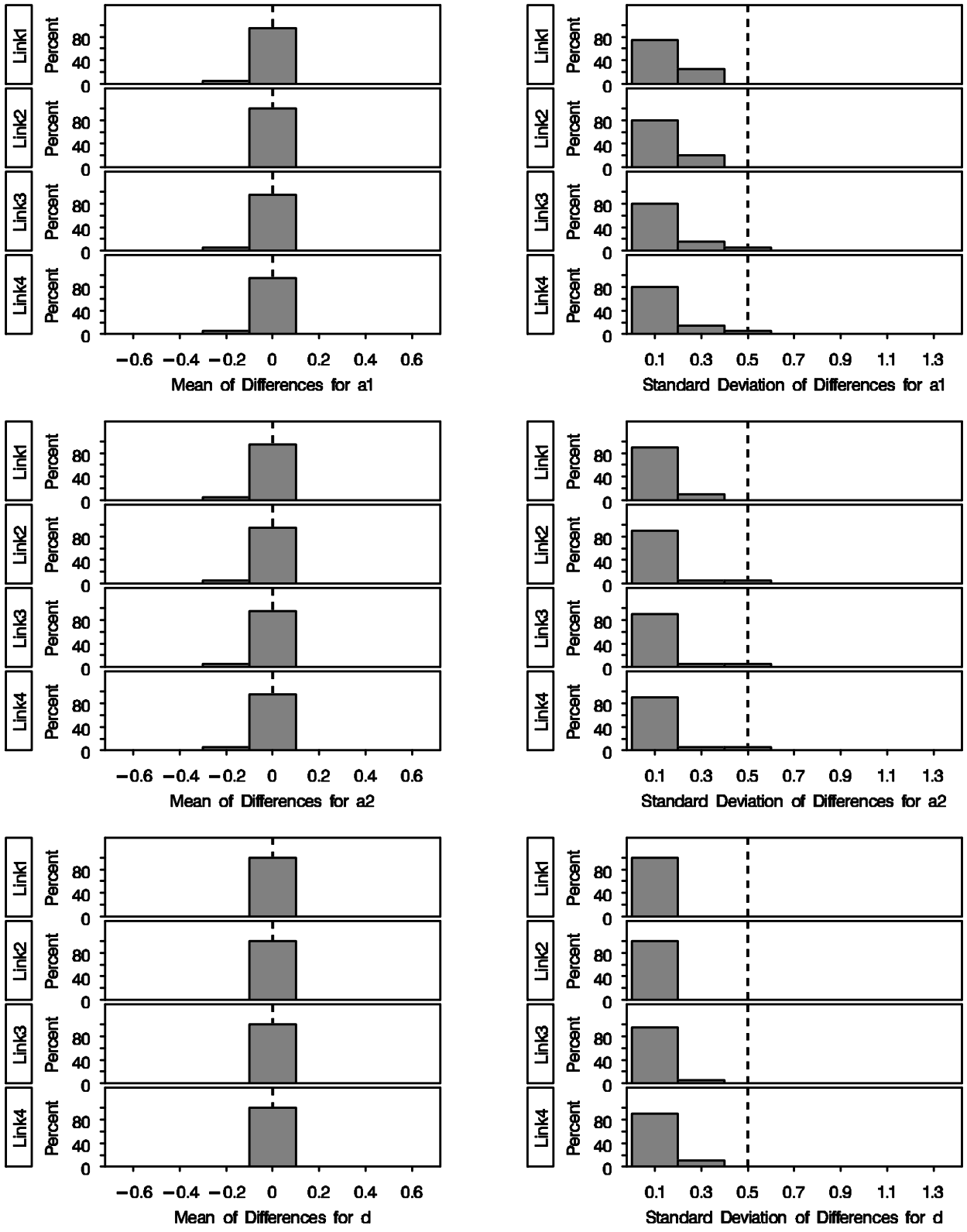


Figure A-11. Accuracy and stability for different linking methods (APP, n=20, N=2000, G3)

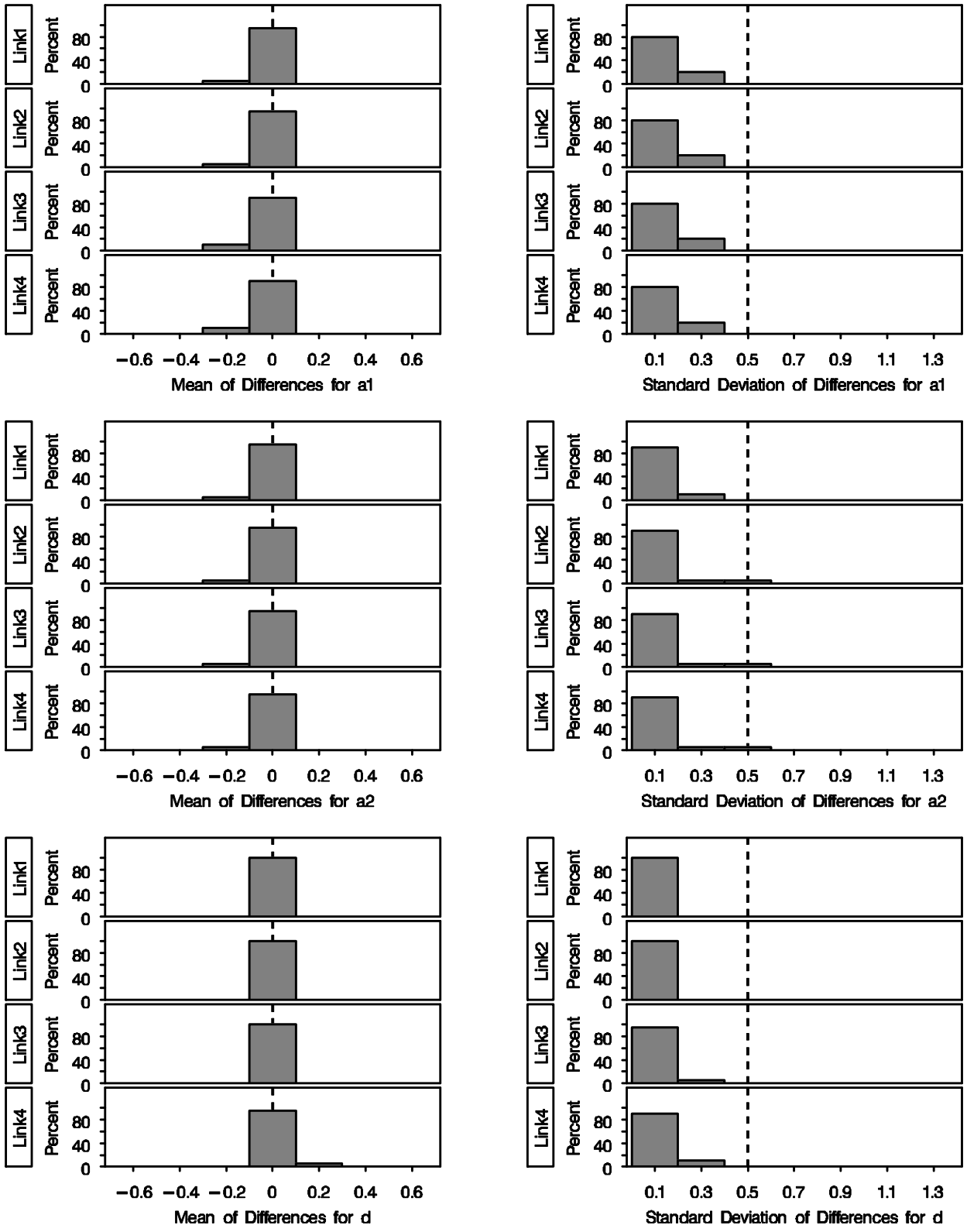


Figure A-12. Accuracy and stability for different linking methods (APP, n=20, N=2000, G4)

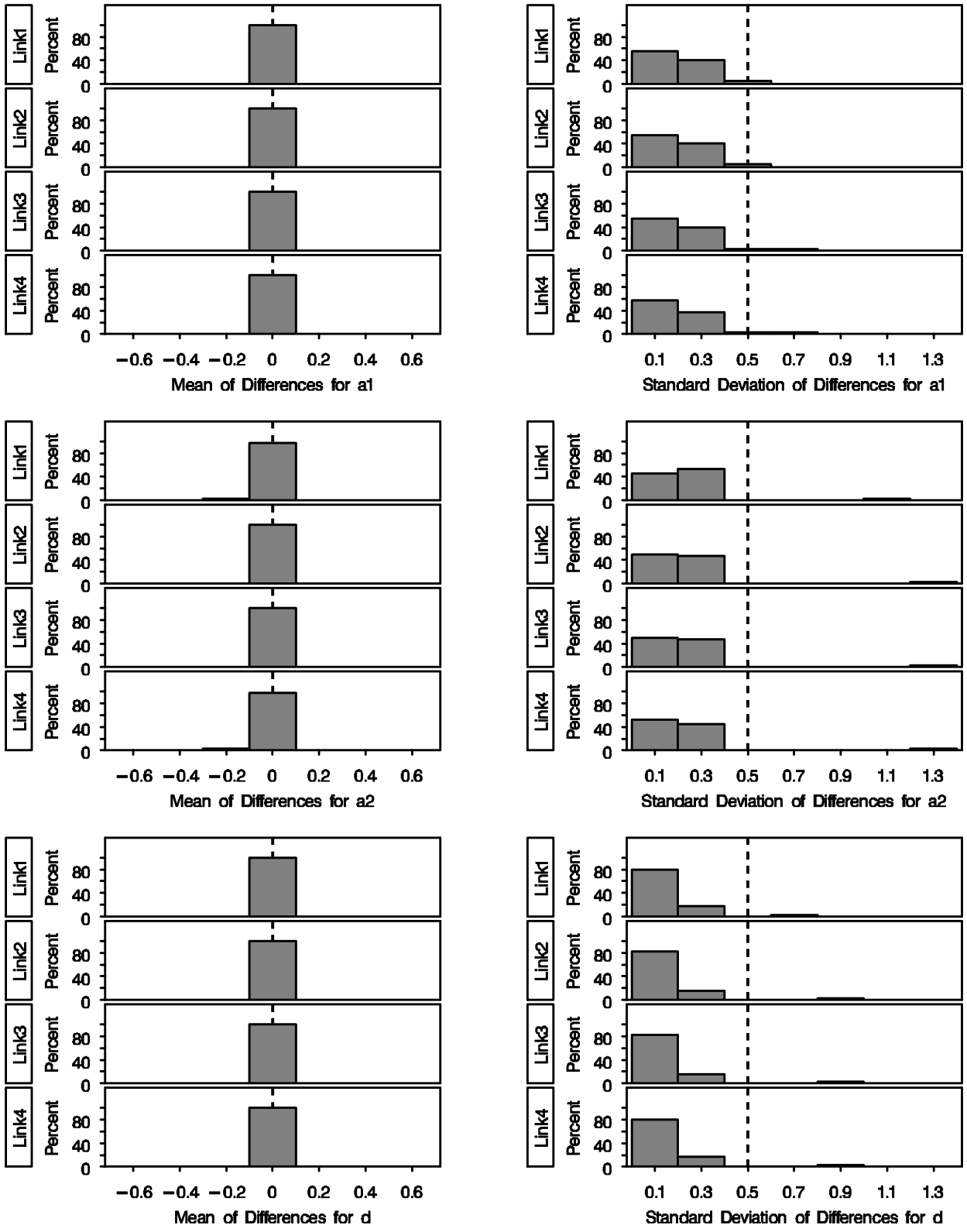


Figure A-13. Accuracy and stability for different linking methods (APP, n=40, N=500, G1)

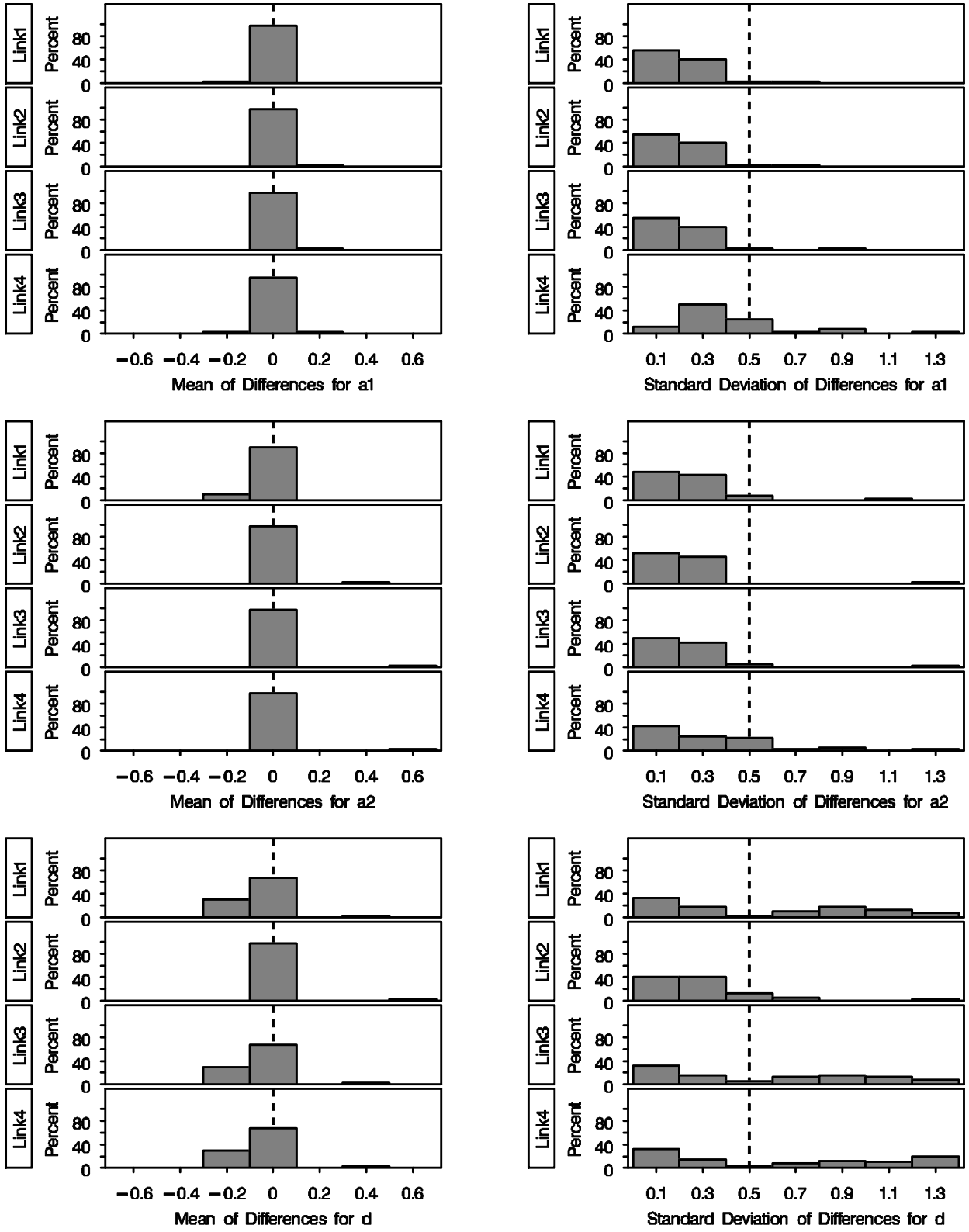


Figure A-14. Accuracy and stability for different linking methods (APP, n=40, N=500, G2)

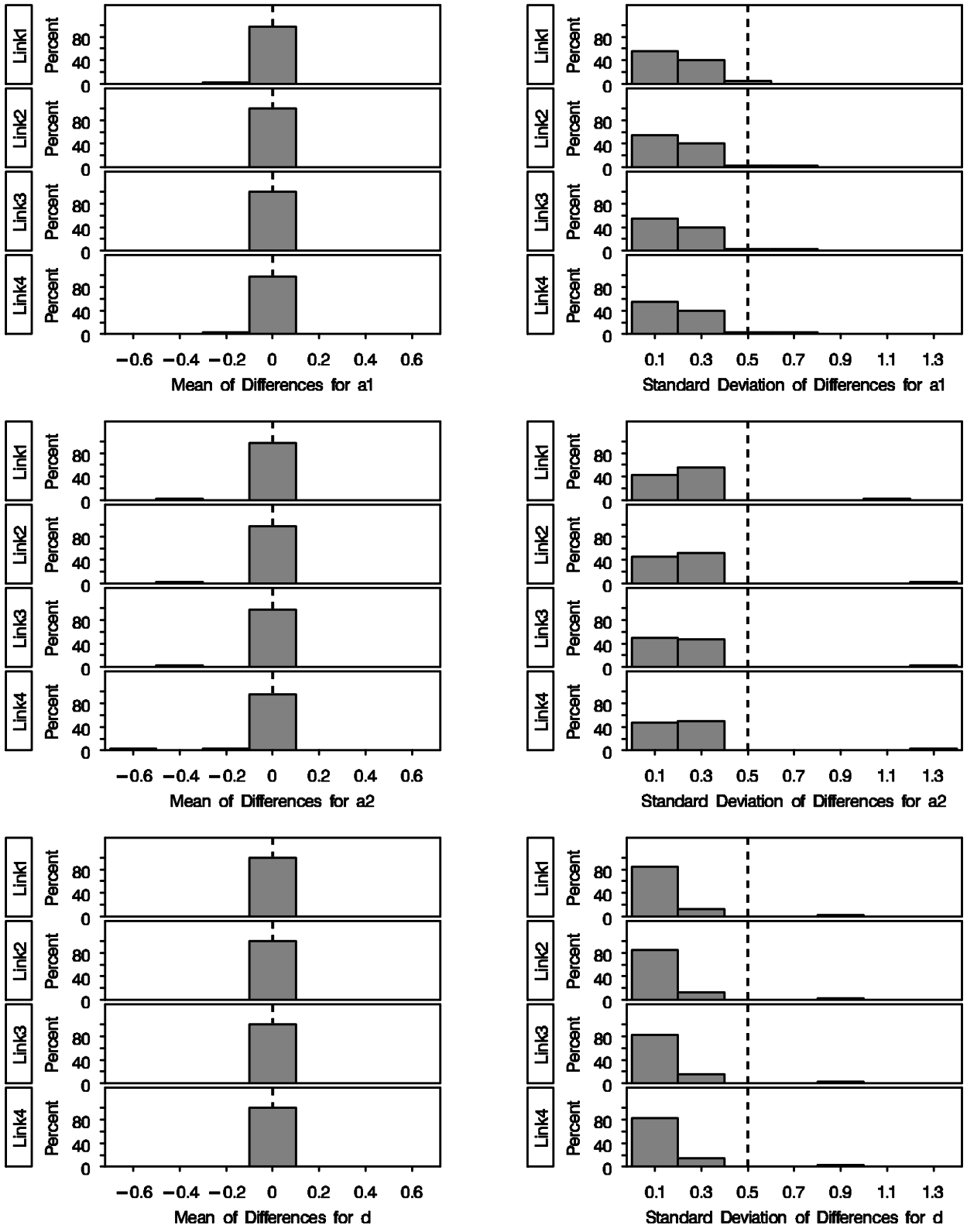


Figure A-15. Accuracy and stability for different linking methods (APP, n=40, N=500, G3)

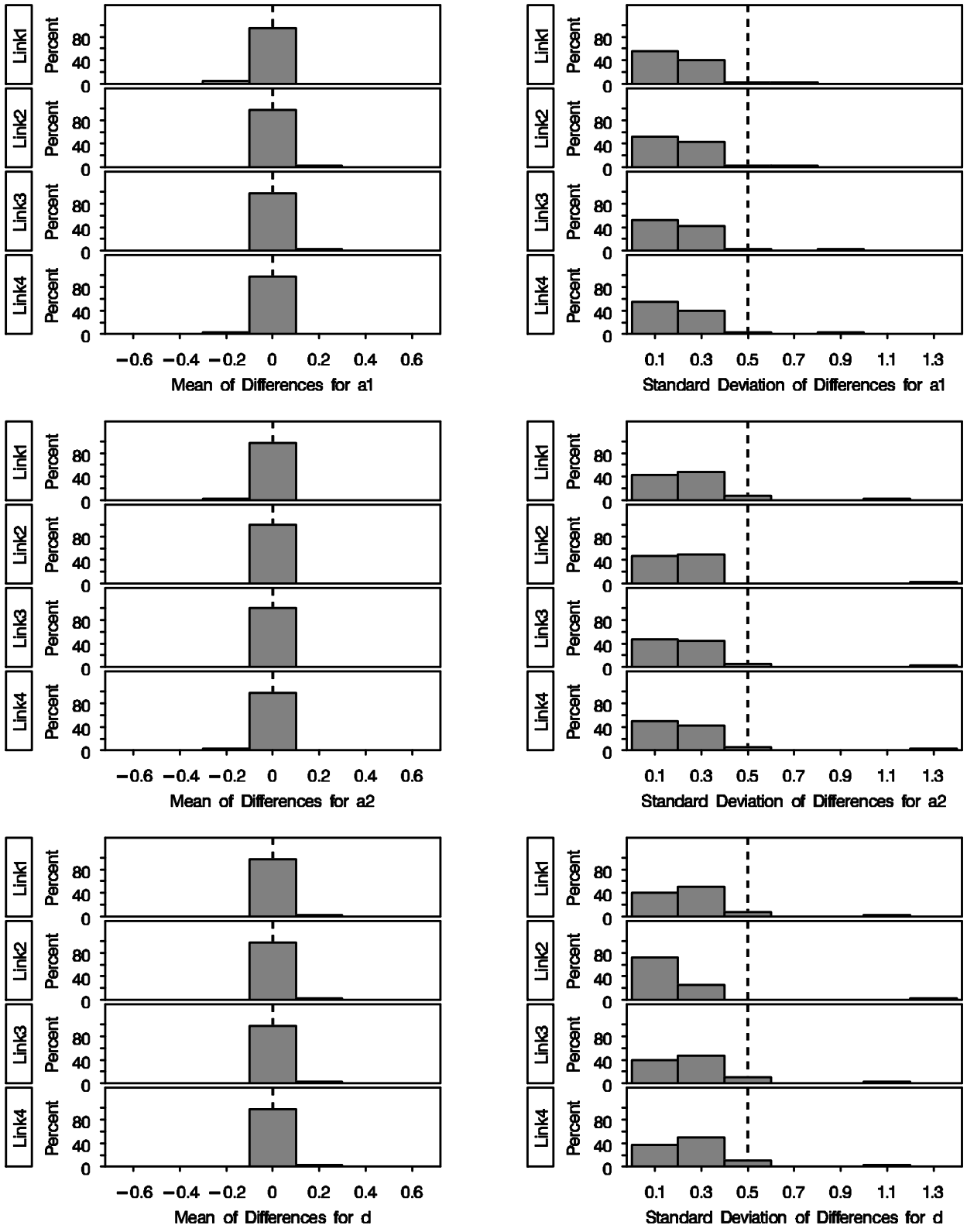


Figure A-16. Accuracy and stability for different linking methods (APP, n=40, N=500, G4)

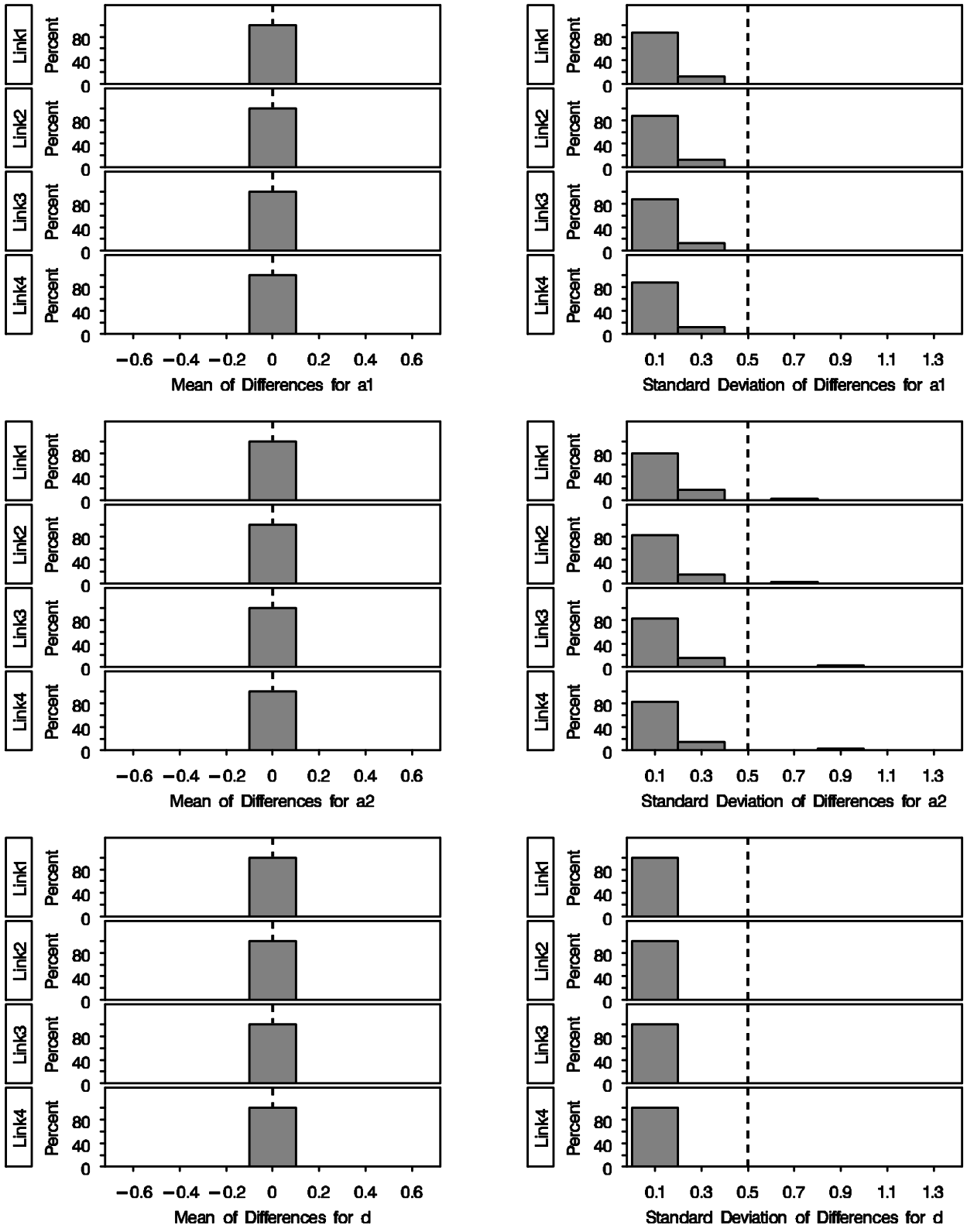


Figure A-17. Accuracy and stability for different linking methods (APP, n=40, N=1000, G1)

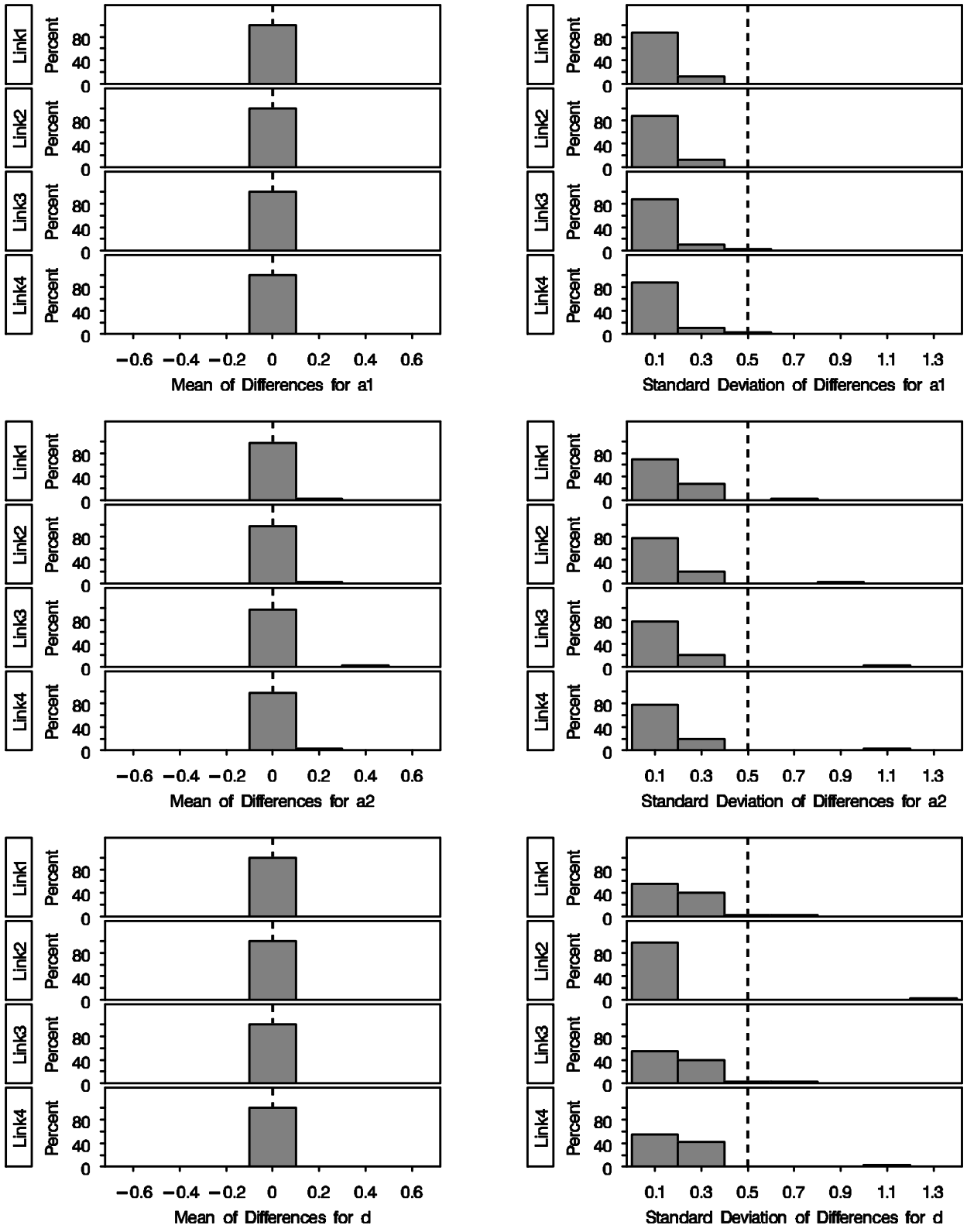


Figure A-18. Accuracy and stability for different linking methods (APP, n=40, N=1000, G2)

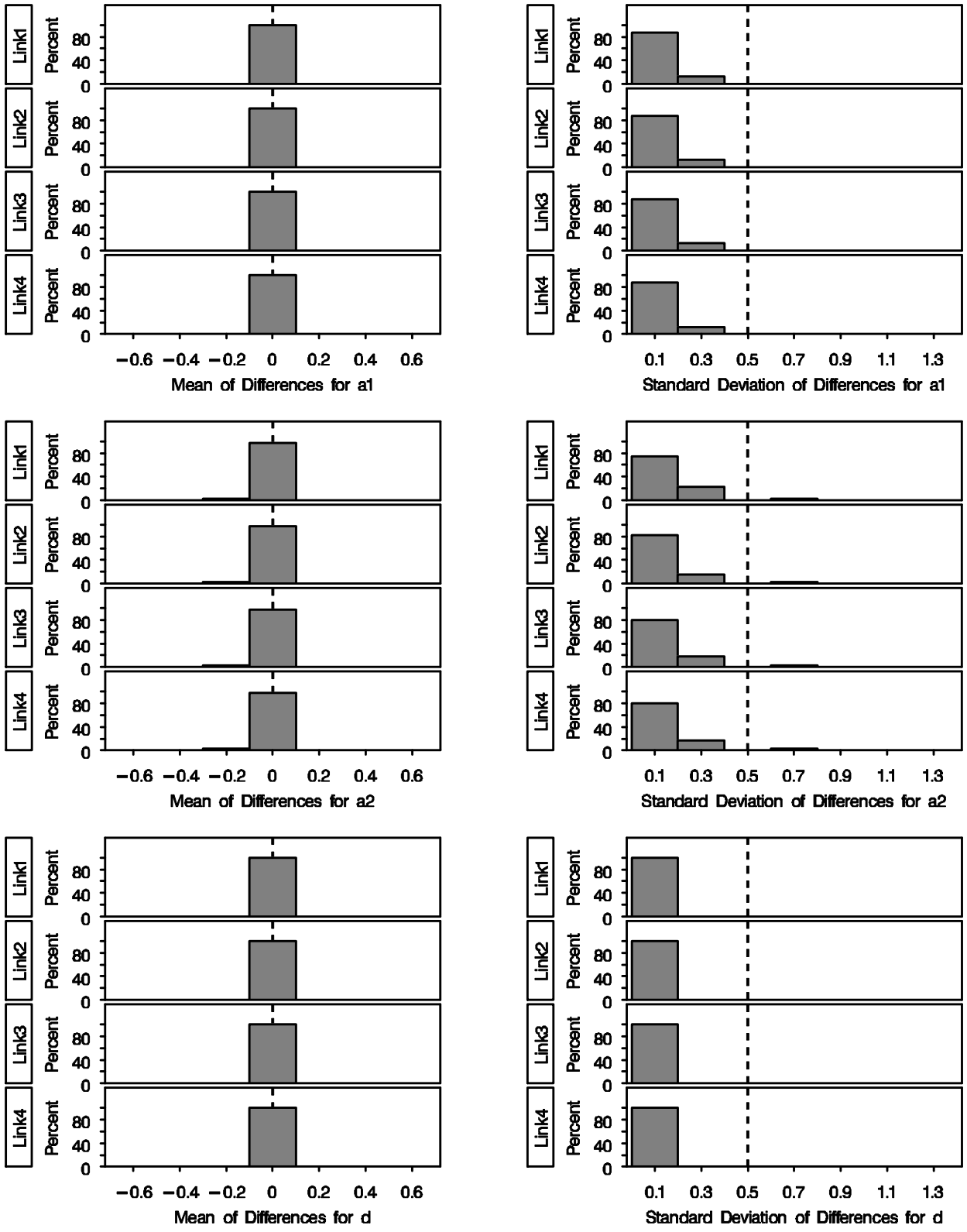


Figure A-19. Accuracy and stability for different linking methods (APP, n=40, N=1000, G3)

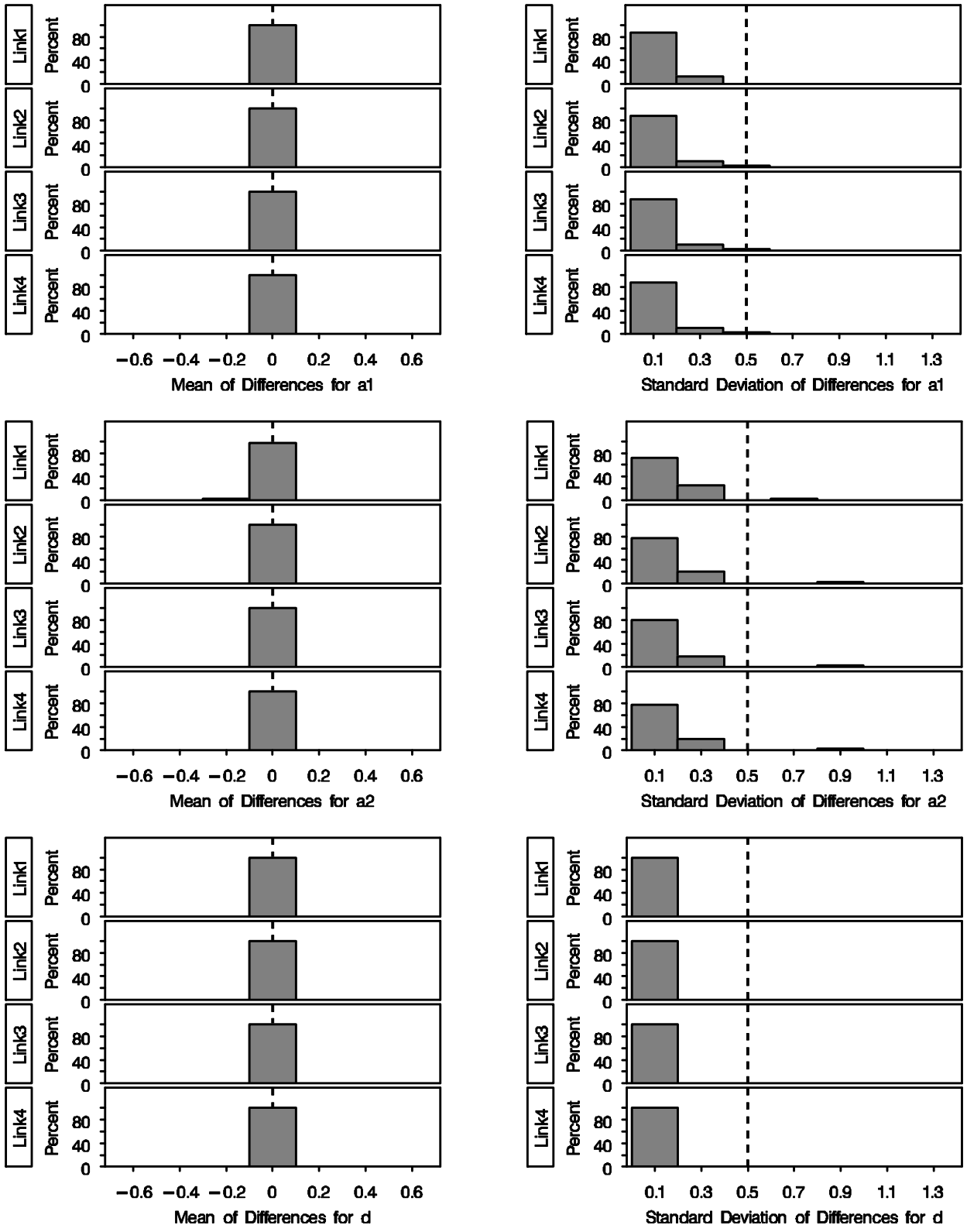


Figure A-20. Accuracy and stability for different linking methods (APP, n=40, N=1000, G4)

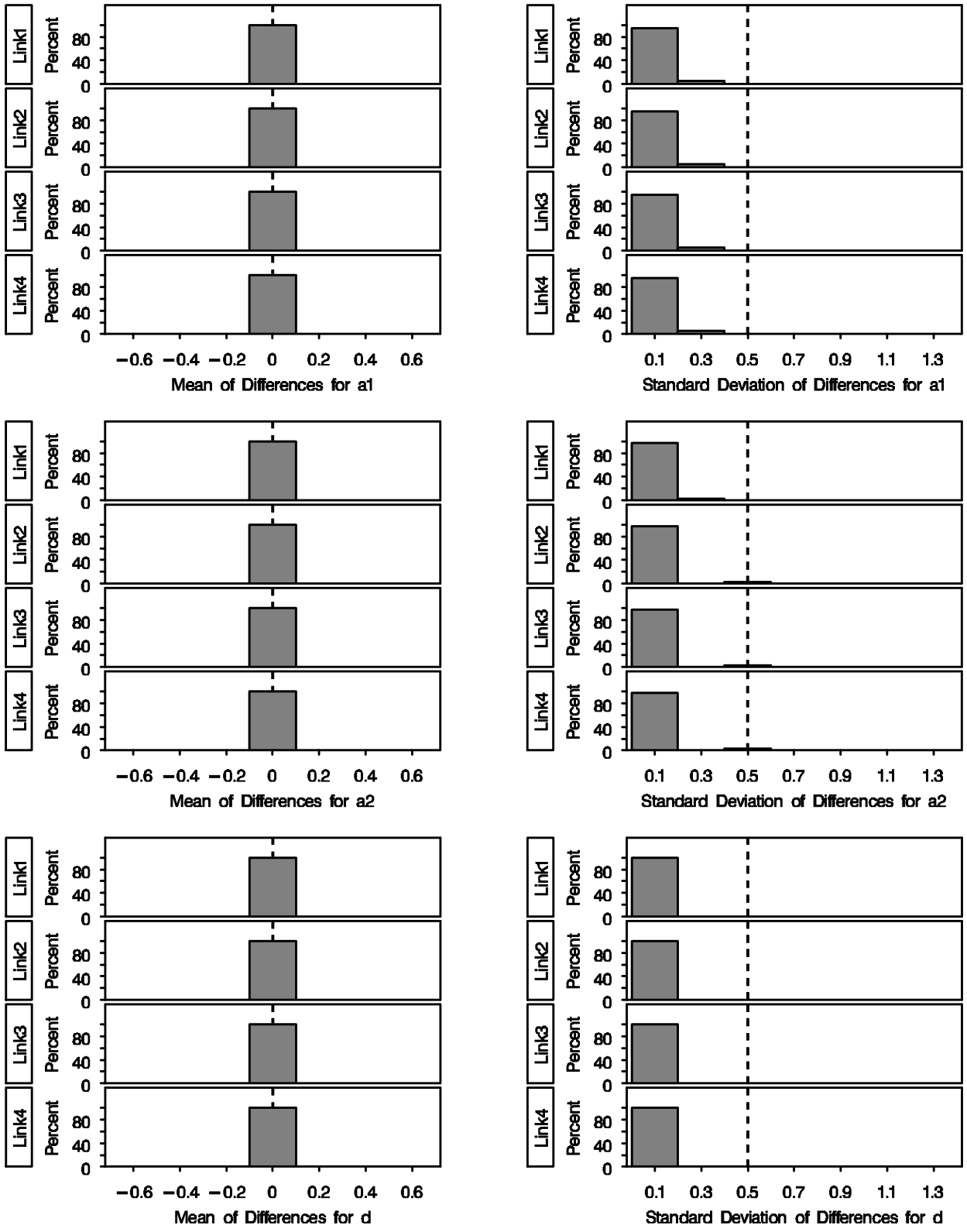


Figure A-21. Accuracy and stability for different linking methods (APP, n=40, N=2000, G1)

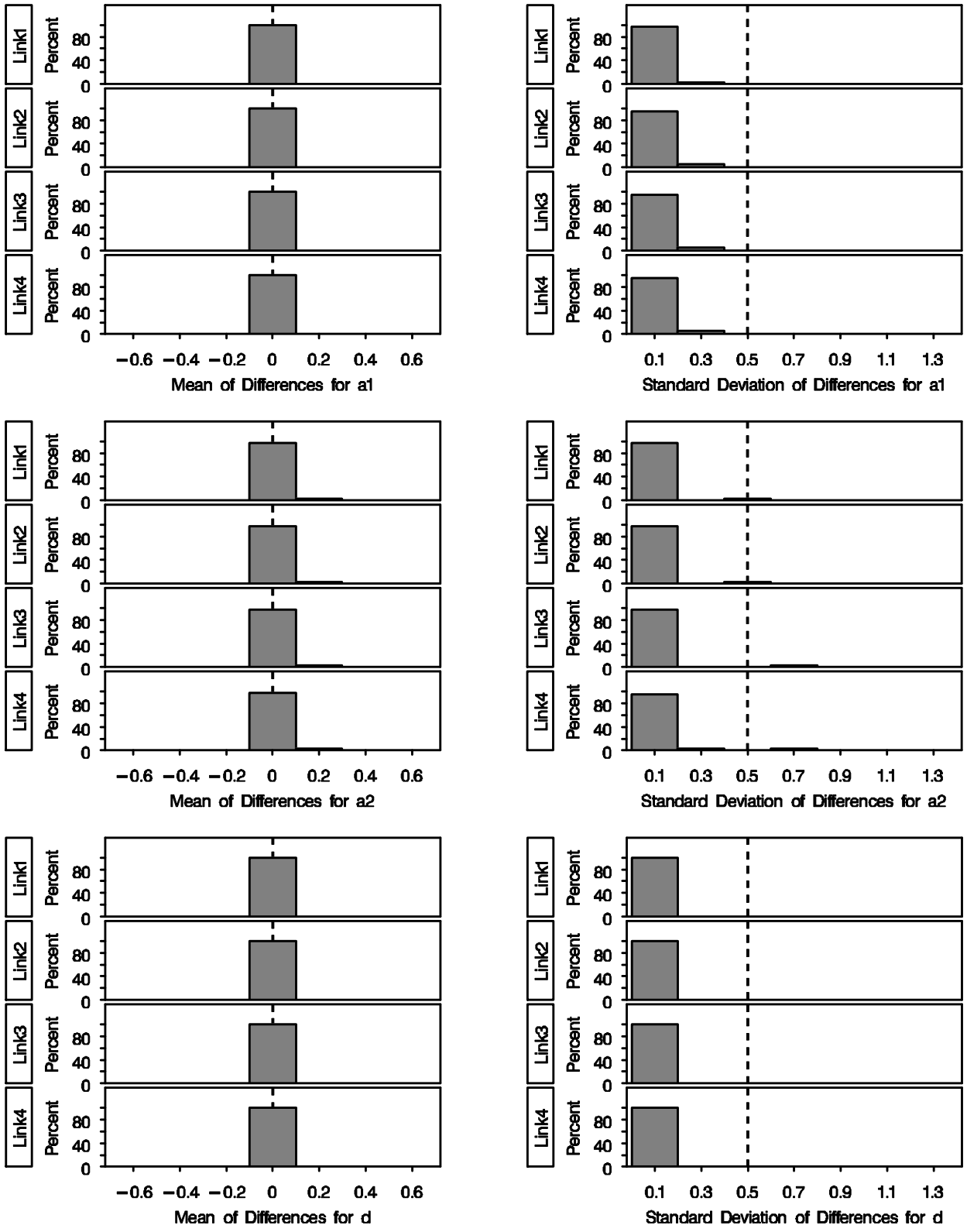


Figure A-22. Accuracy and stability for different linking methods (APP, n=40, N=2000, G2)

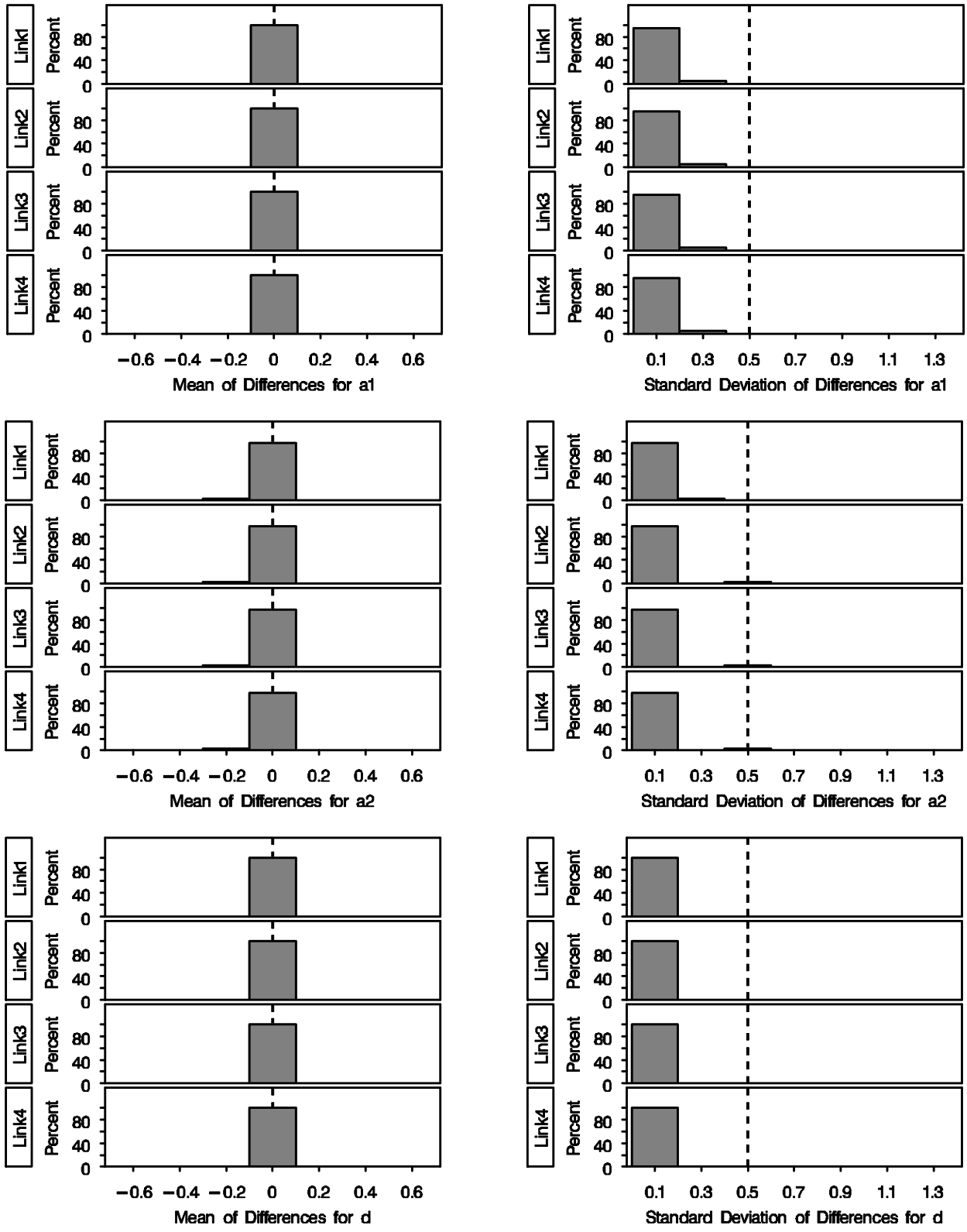


Figure A-23. Accuracy and stability for different linking methods (APP, n=40, N=2000, G3)

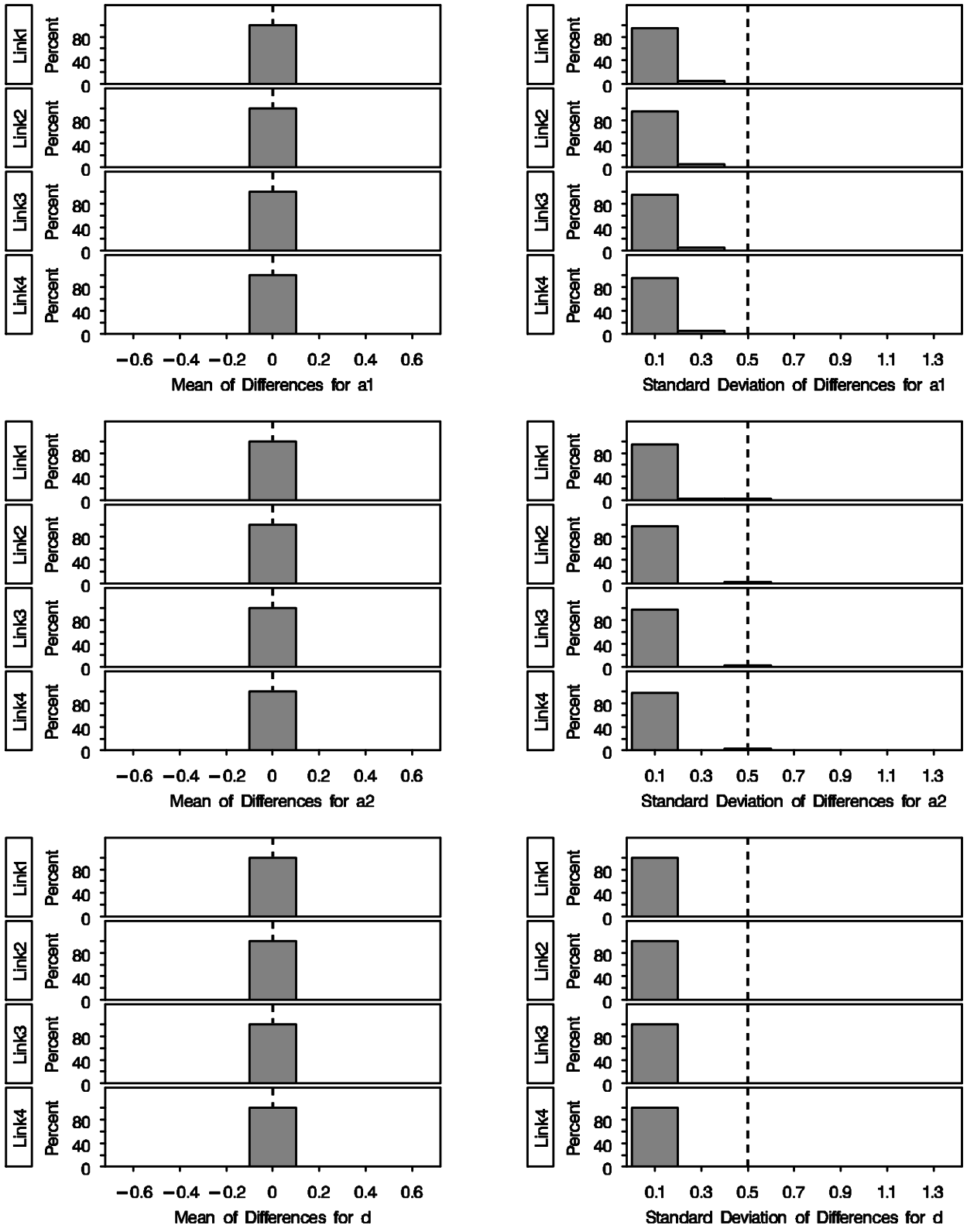


Figure A-24. Accuracy and stability for different linking methods (APP, n=40, N=2000, G4)

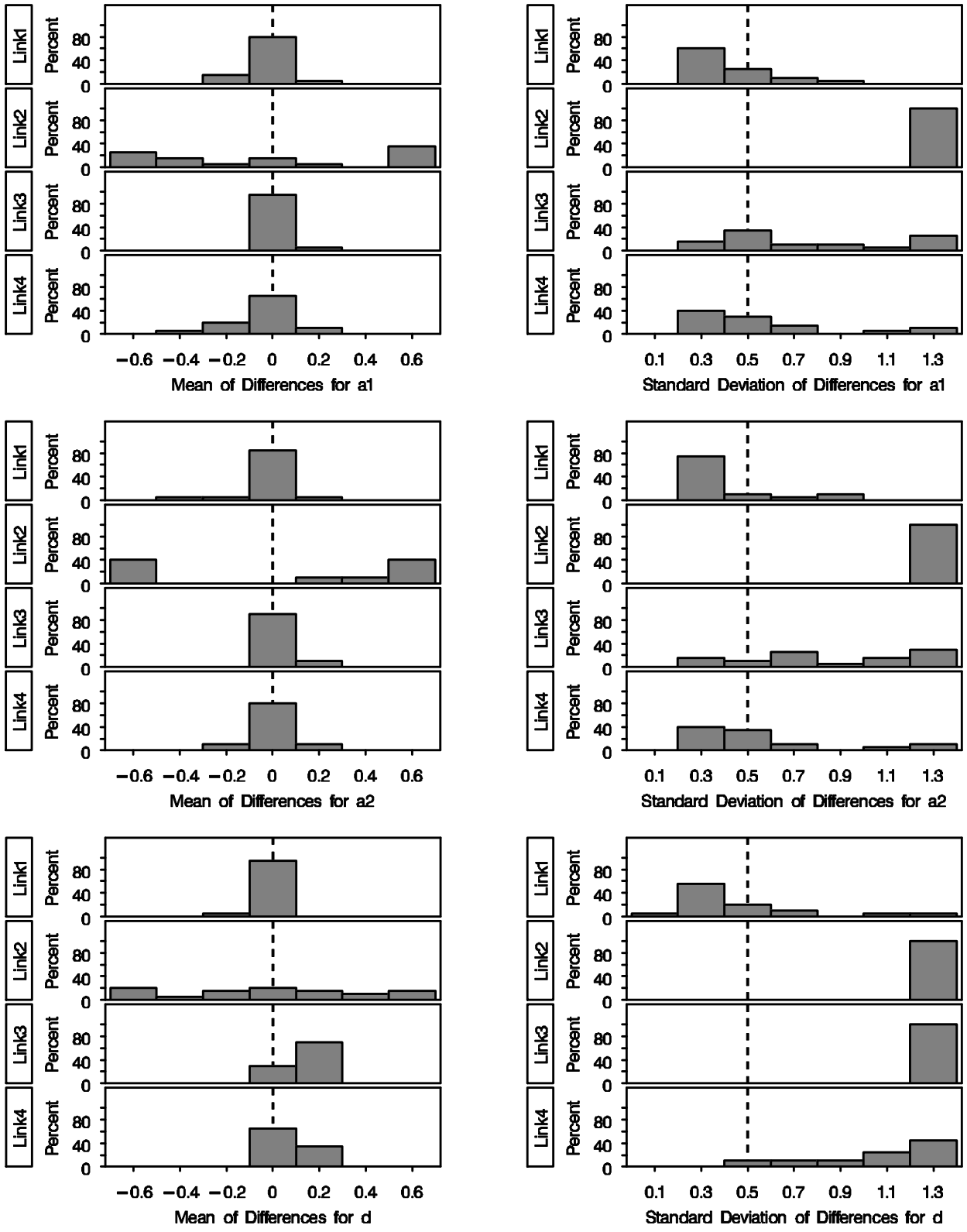


Figure A-25. Accuracy and stability for different linking methods (COM, n=20, N=500, G1)

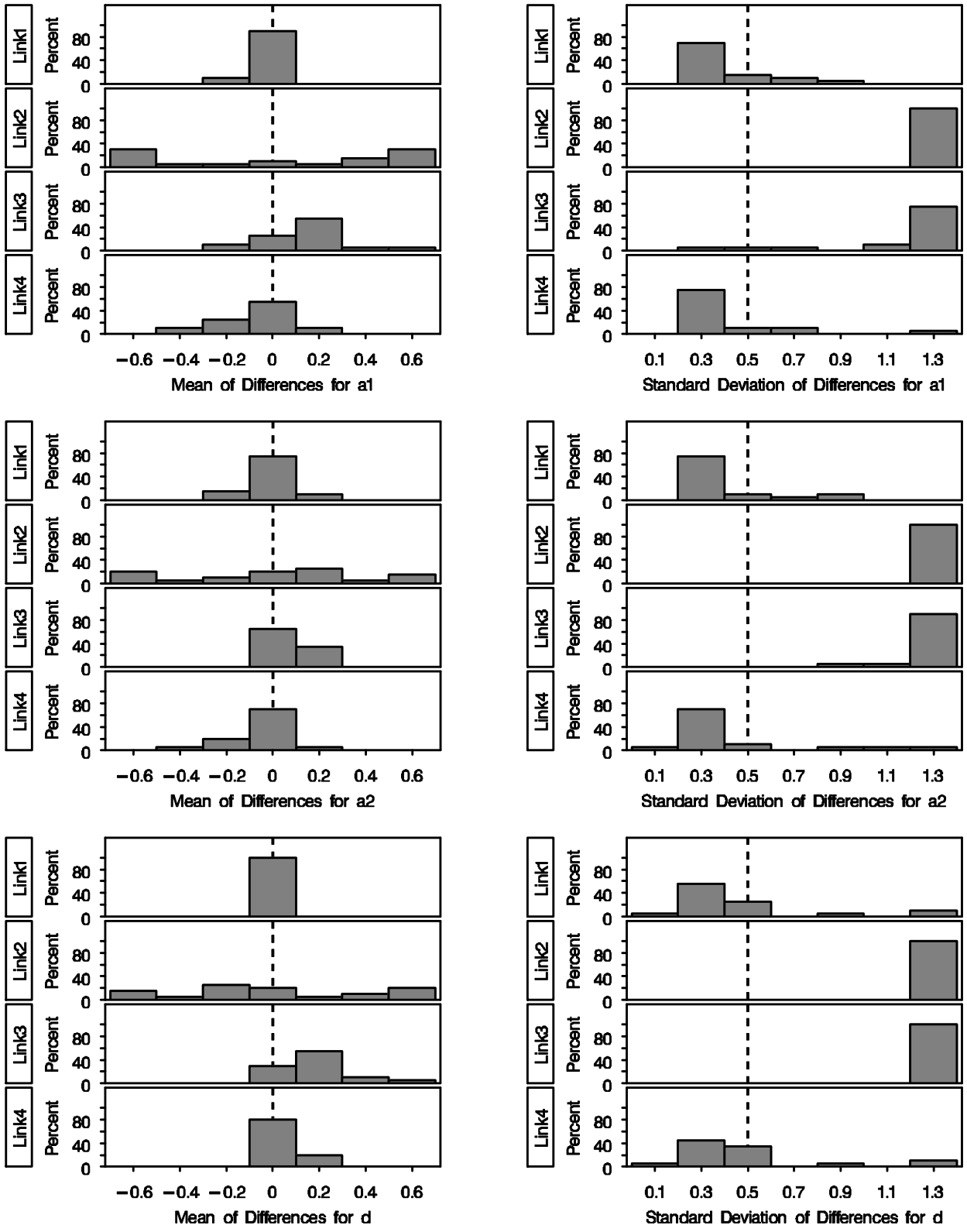


Figure A-26. Accuracy and stability for different linking methods (COM, n=20, N=500, G2)

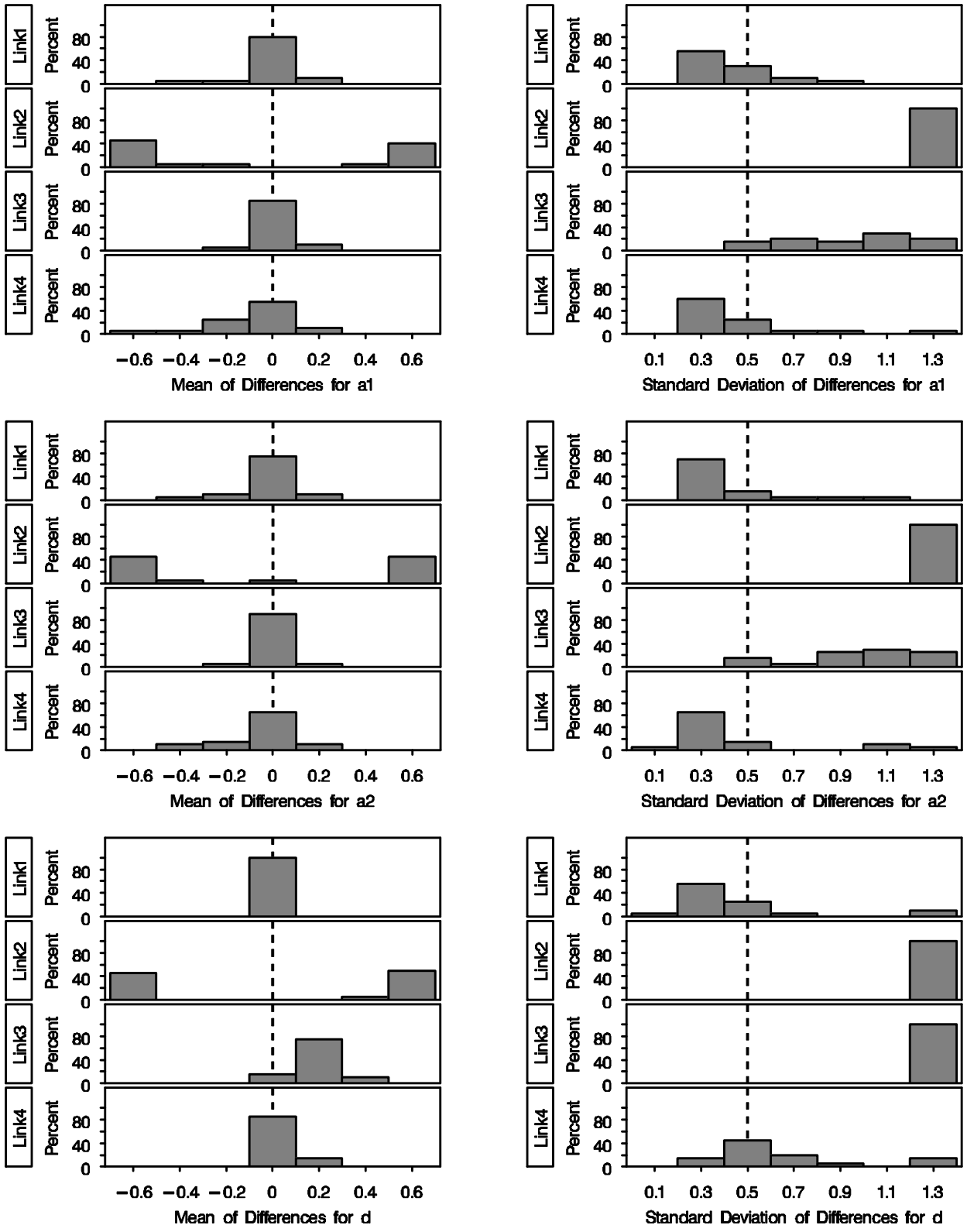


Figure A-27. Accuracy and stability for different linking methods (COM, n=20, N=500, G3)

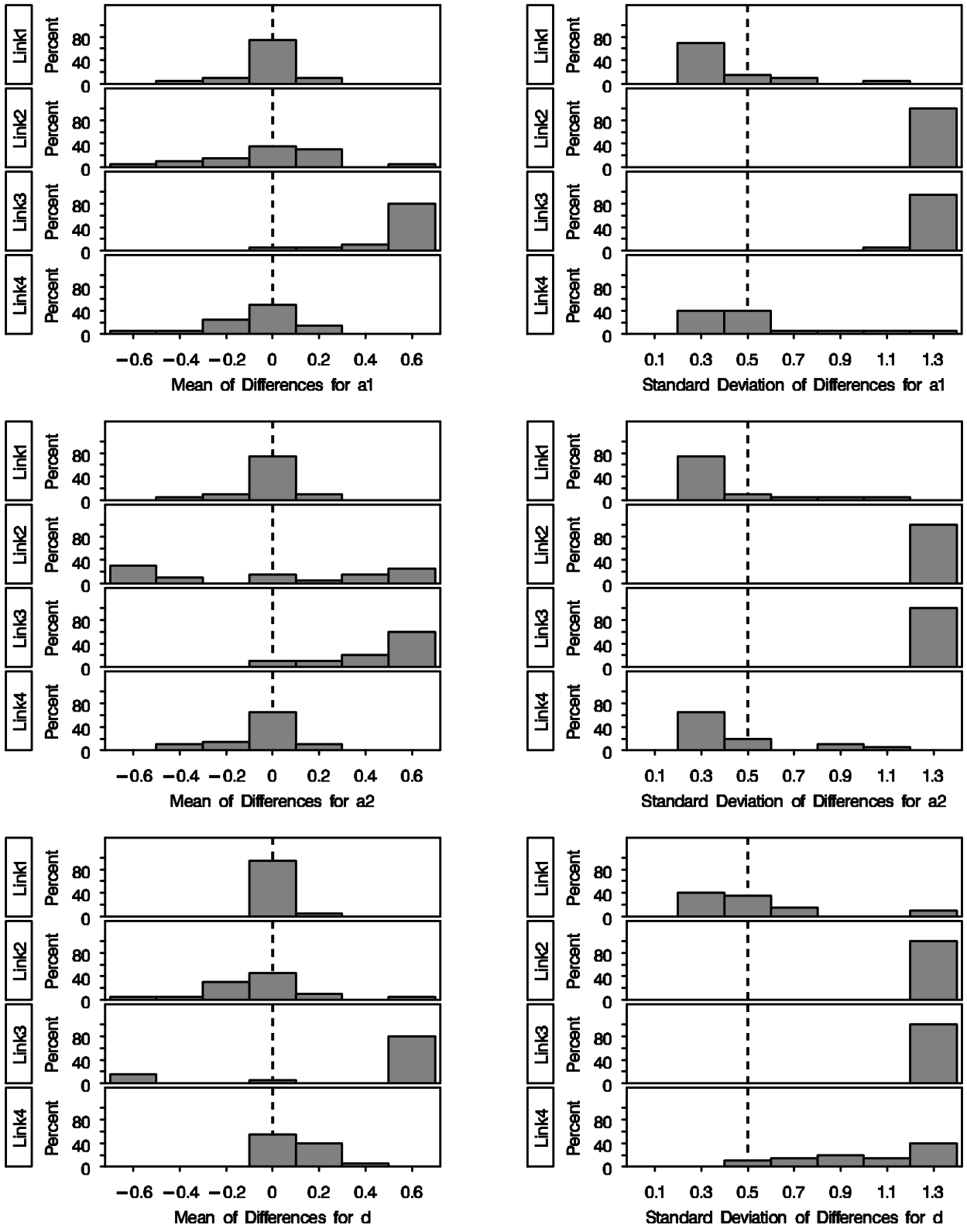


Figure A-28. Accuracy and stability for different linking methods (COM, n=20, N=500, G4)

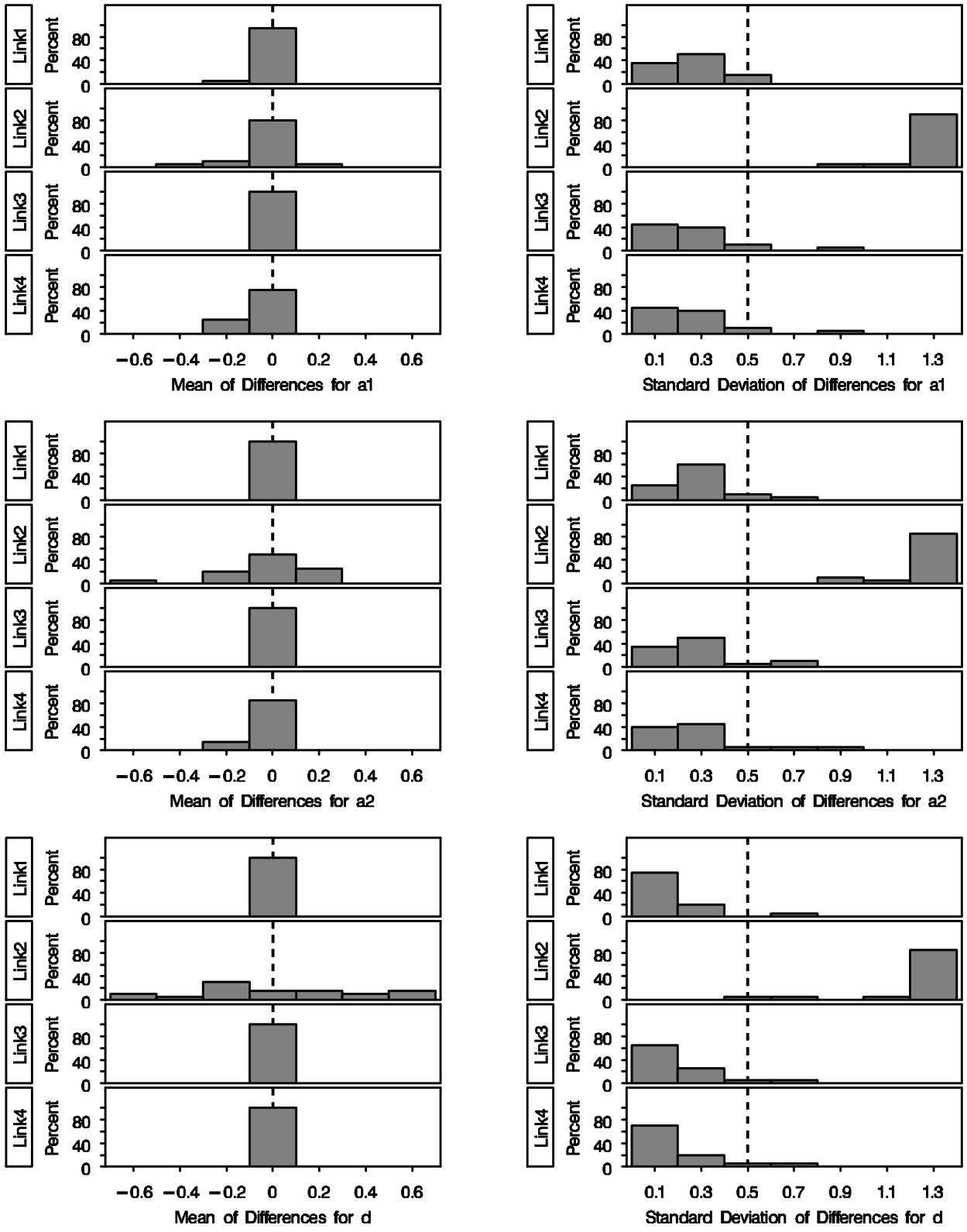


Figure A-29. Accuracy and stability for different linking methods (COM, n=20, N=1000, G1)

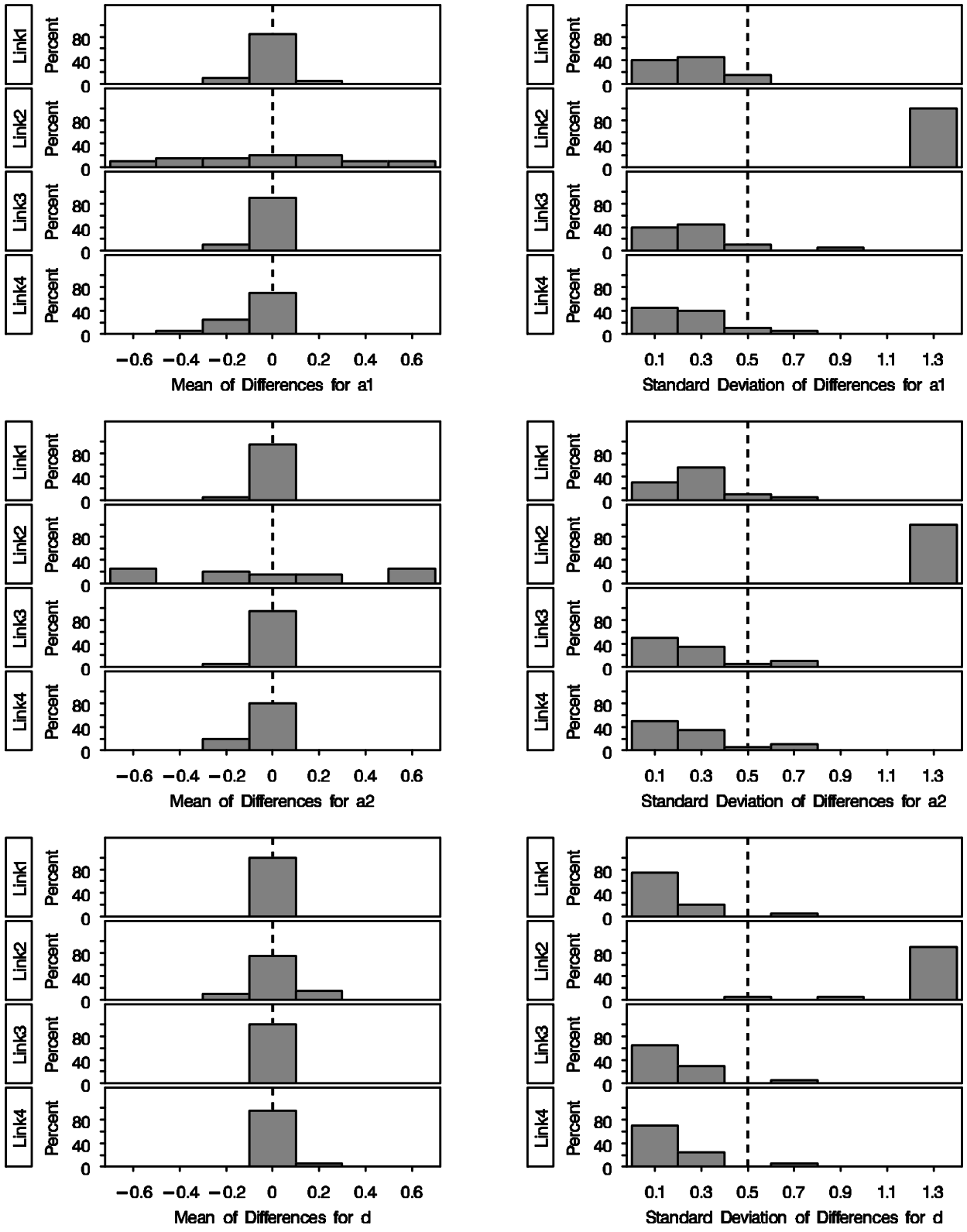


Figure A-30. Accuracy and stability for different linking methods (COM, n=20, N=1000, G2)

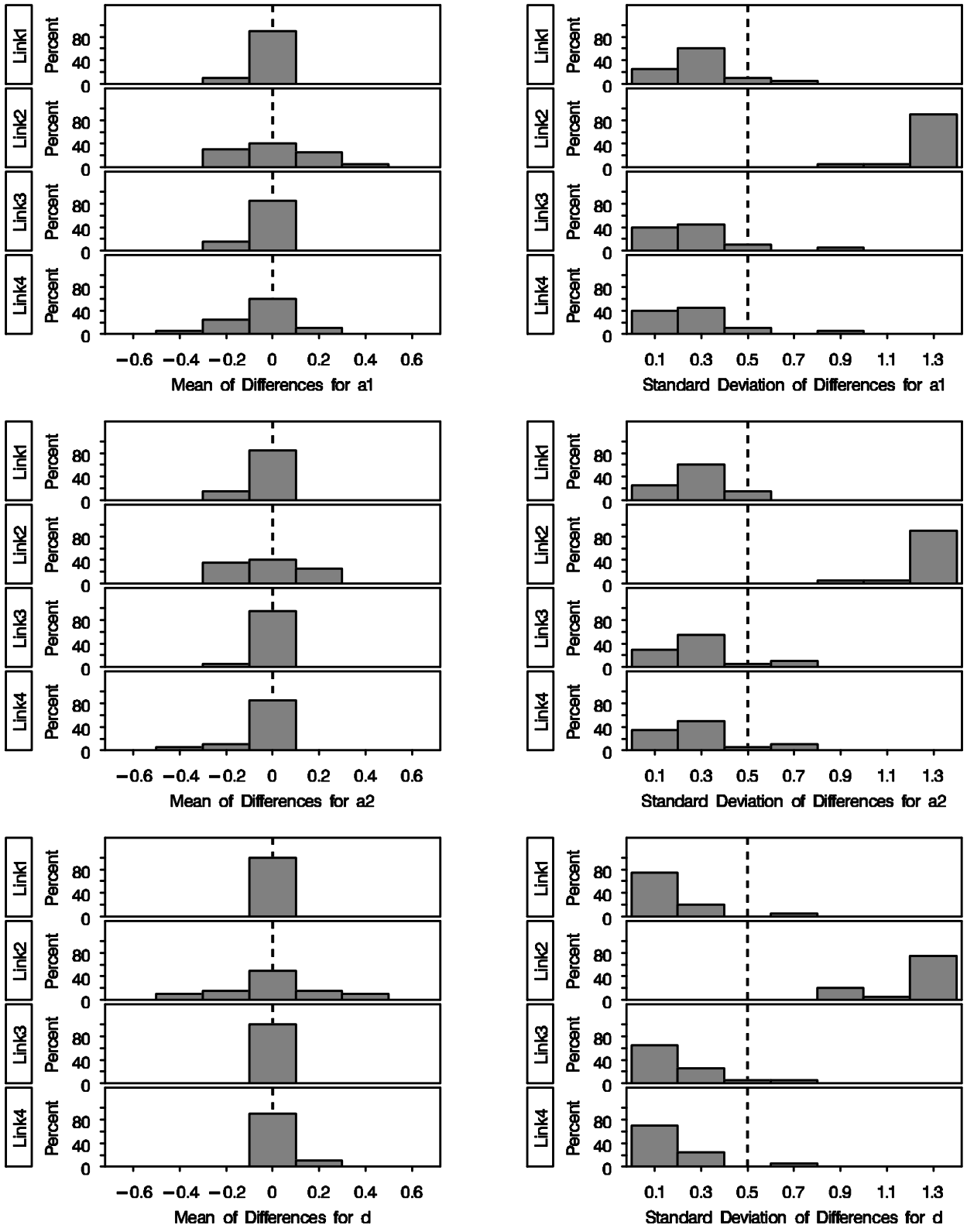


Figure A-31. Accuracy and stability for different linking methods (COM, n=20, N=1000, G3)

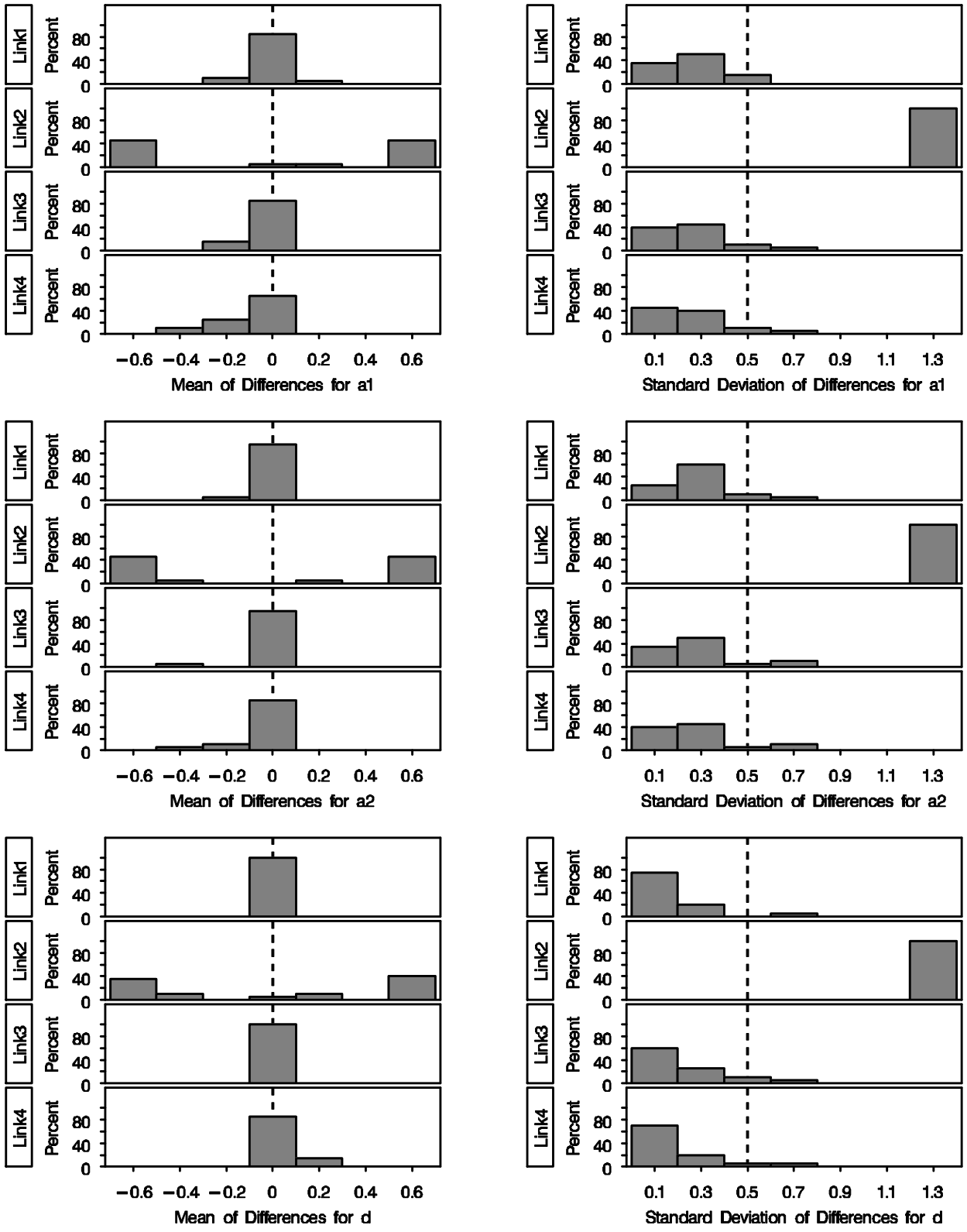


Figure A-32. Accuracy and stability for different linking methods (COM, n=20, N=1000, G4)

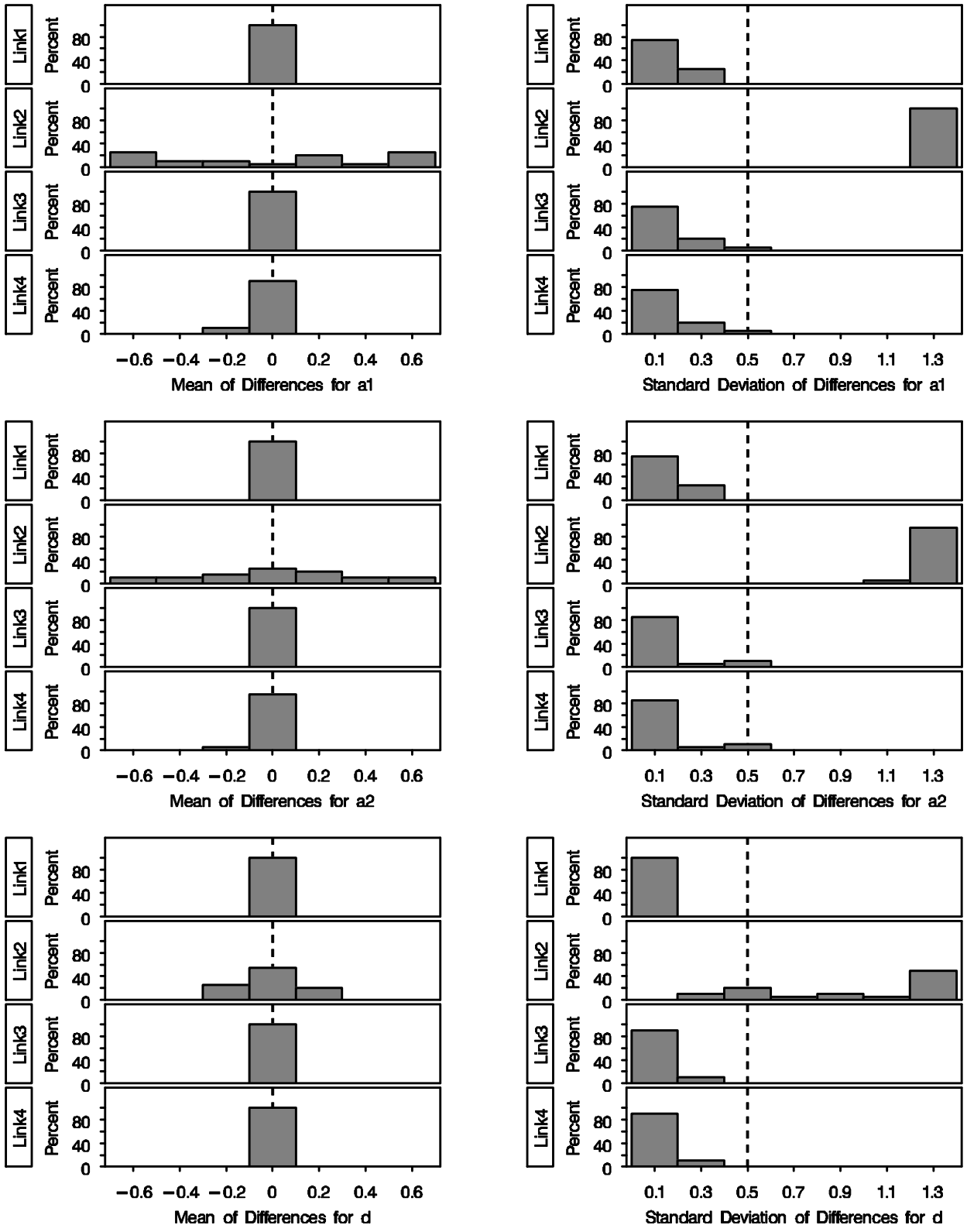


Figure A-33. Accuracy and stability for different linking methods (COM, n=20, N=2000, G1)

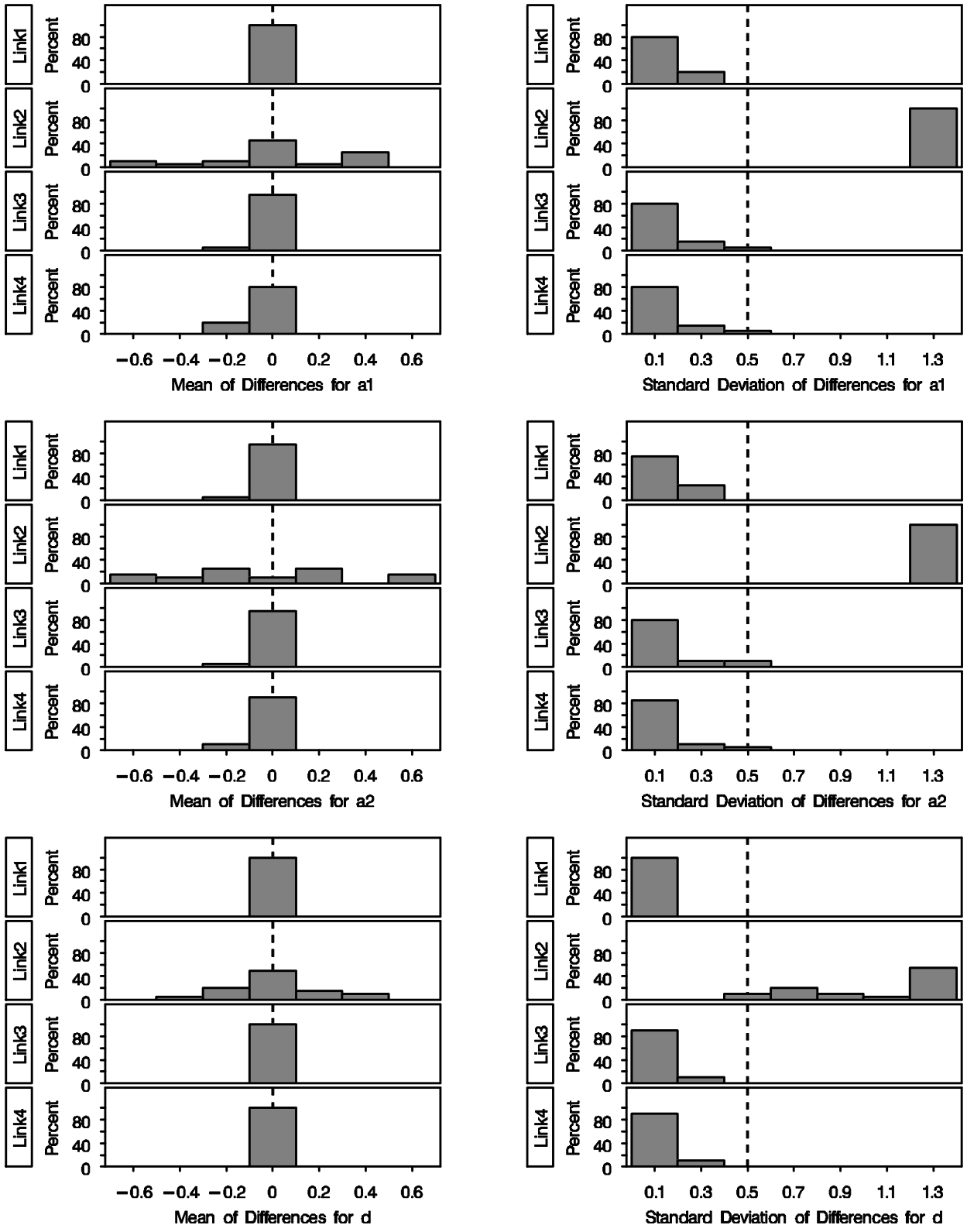


Figure A-34. Accuracy and stability for different linking methods (COM, n=20, N=2000, G2)

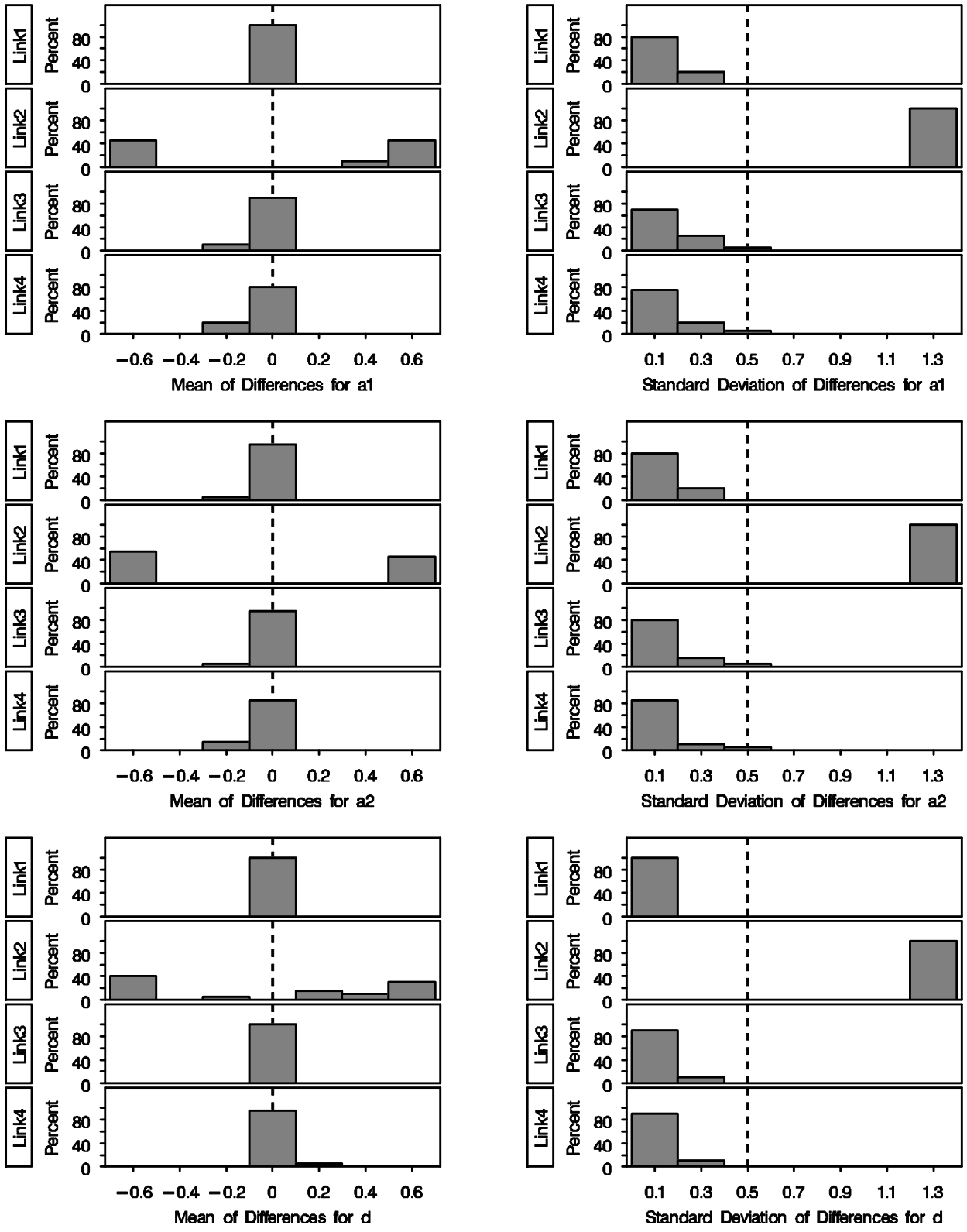


Figure A-35. Accuracy and stability for different linking methods (COM, n=20, N=2000, G3)

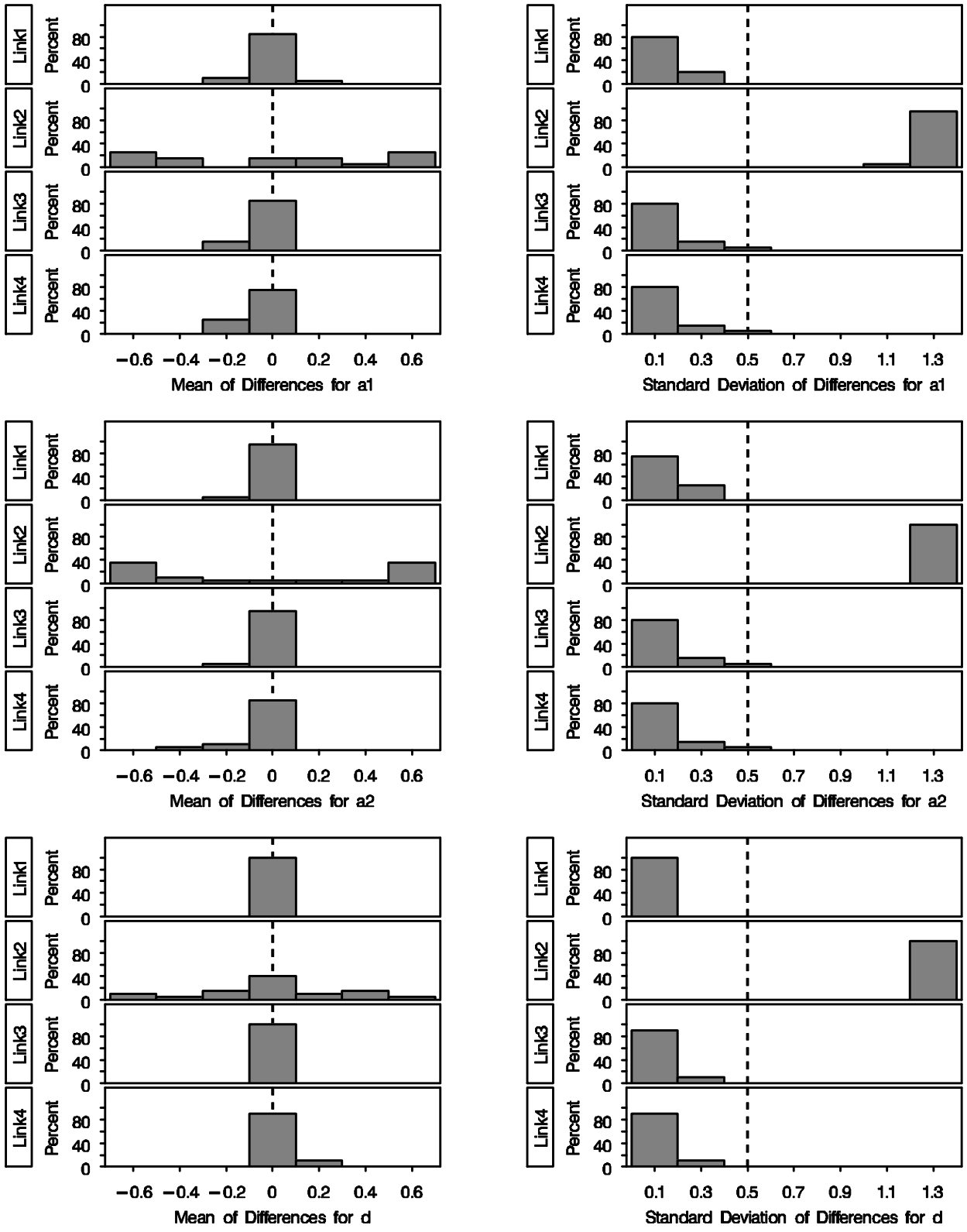


Figure A-36. Accuracy and stability for different linking methods (COM, n=20, N=2000, G4)

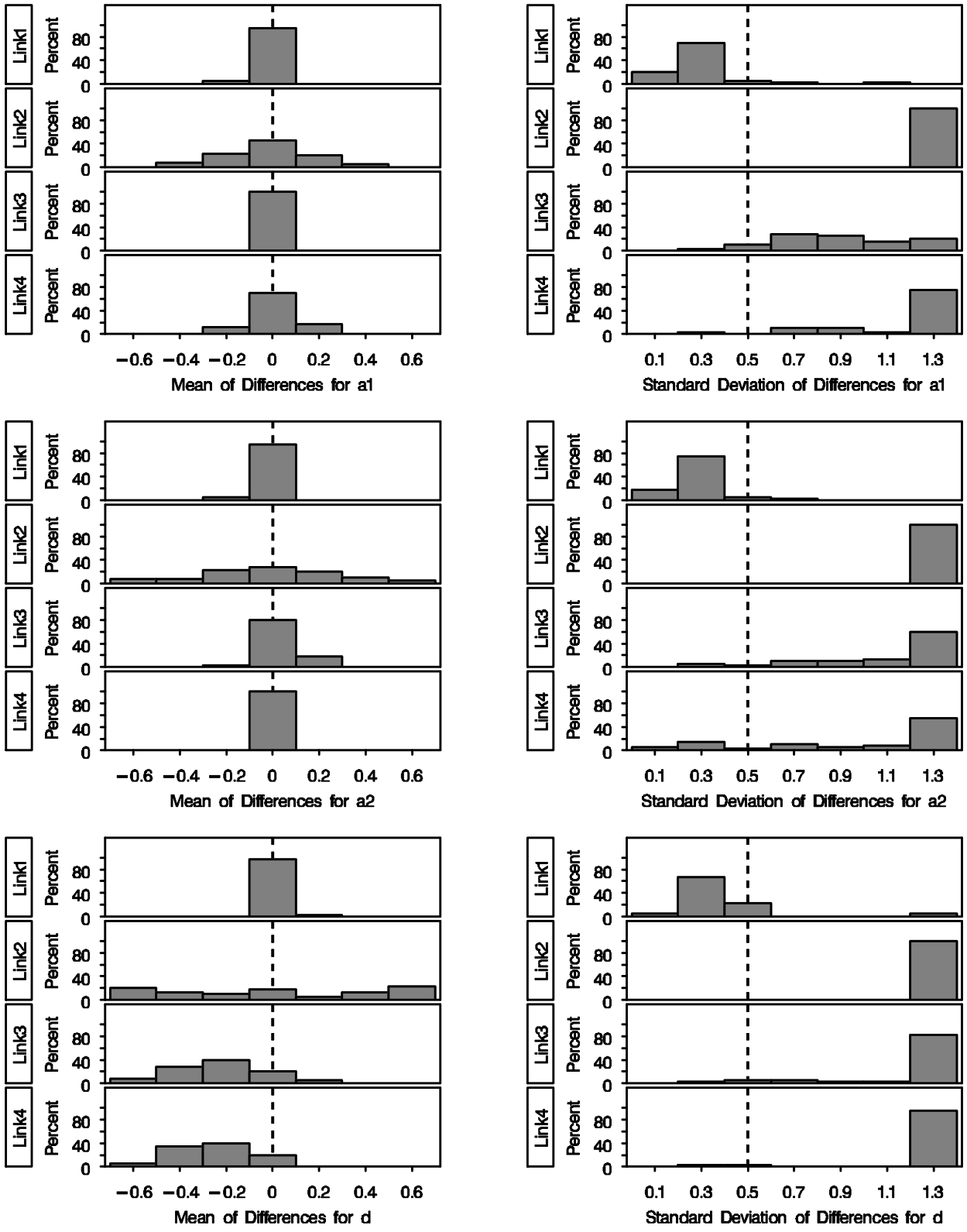


Figure A-37. Accuracy and stability for different linking methods (COM, n=40, N=500, G1)

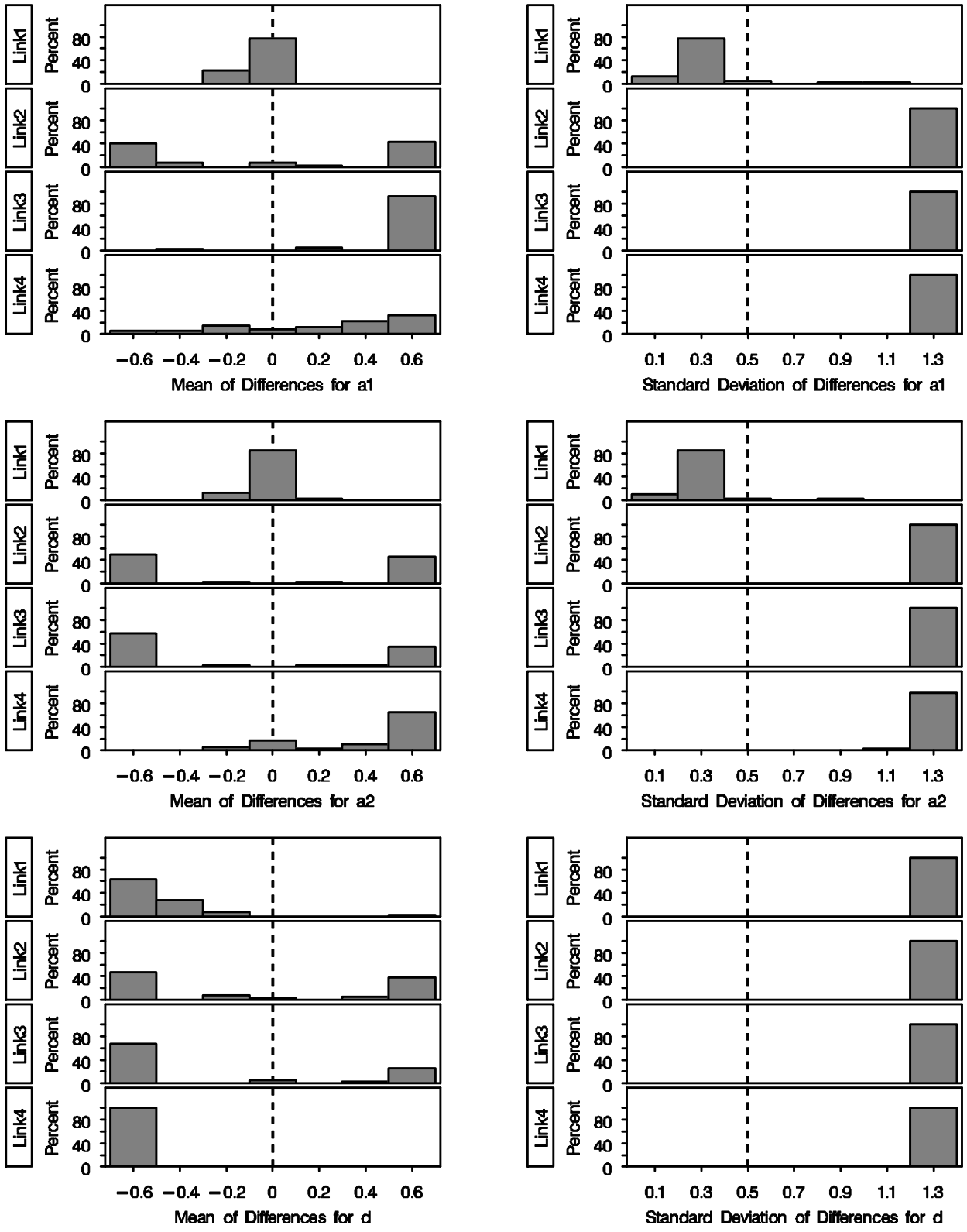


Figure A-38. Accuracy and stability for different linking methods (COM, n=40, N=500, G2)

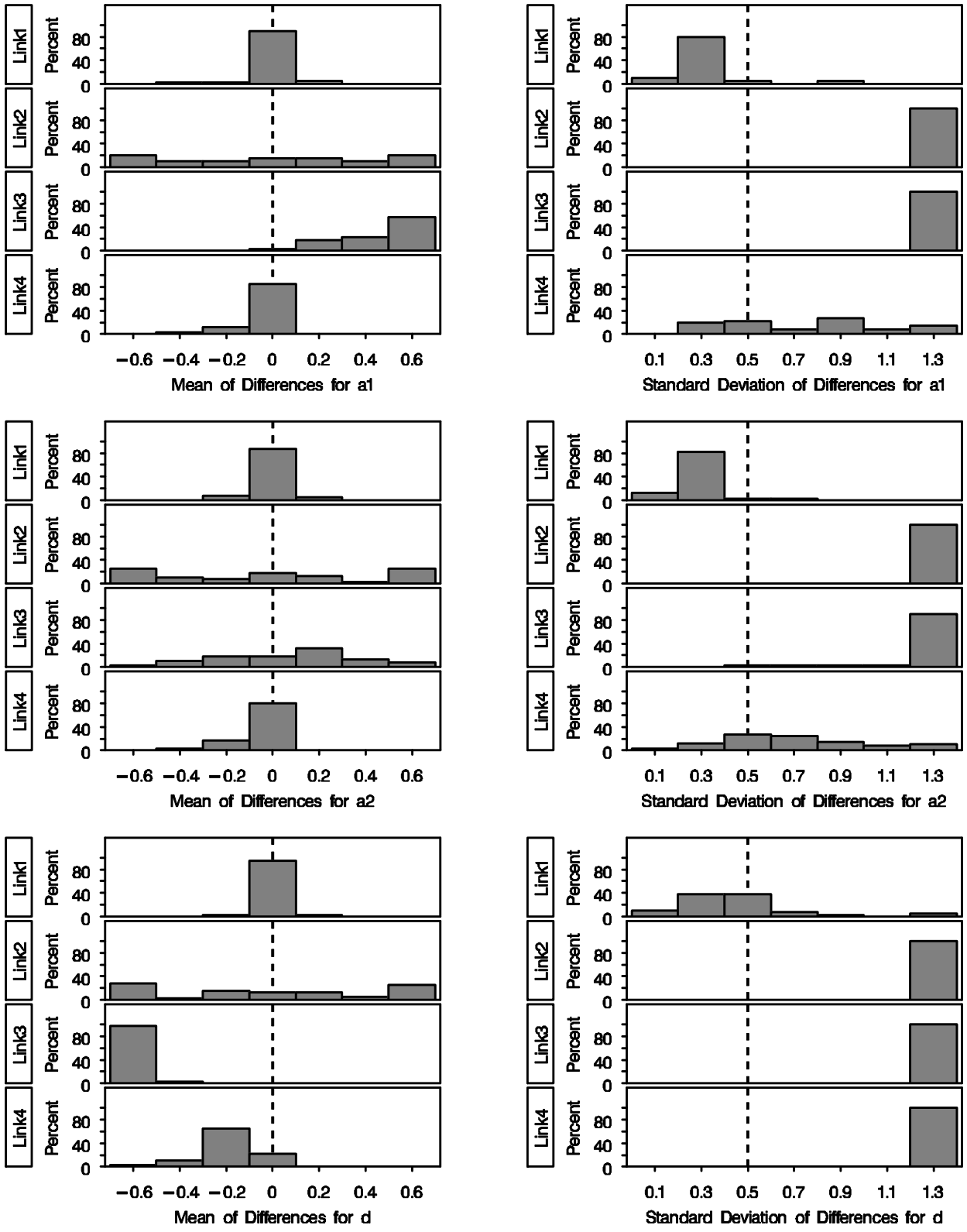


Figure A-39. Accuracy and stability for different linking methods (COM, n=40, N=500, G3)

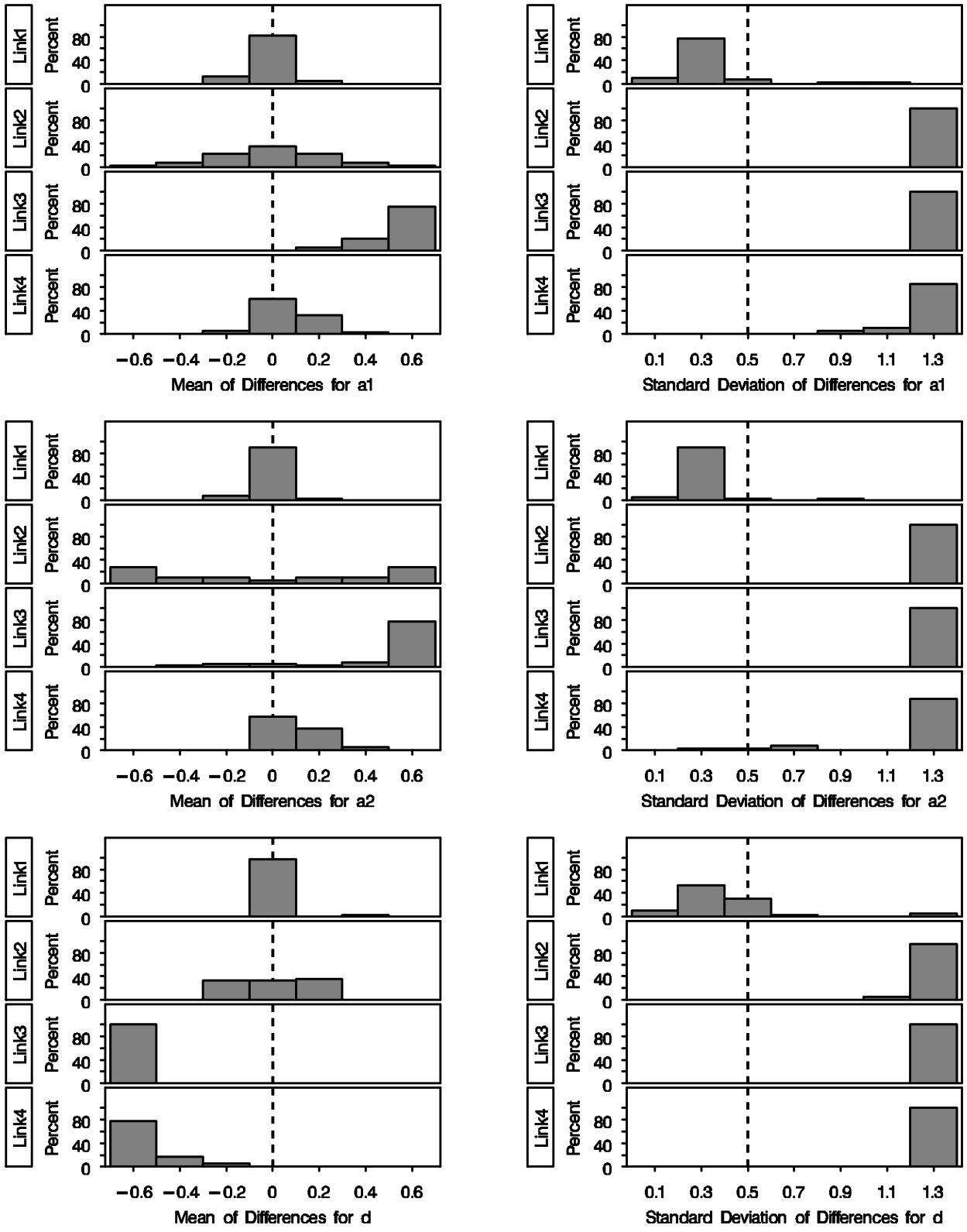


Figure A-40. Accuracy and stability for different linking methods (COM, n=40, N=500, G4)

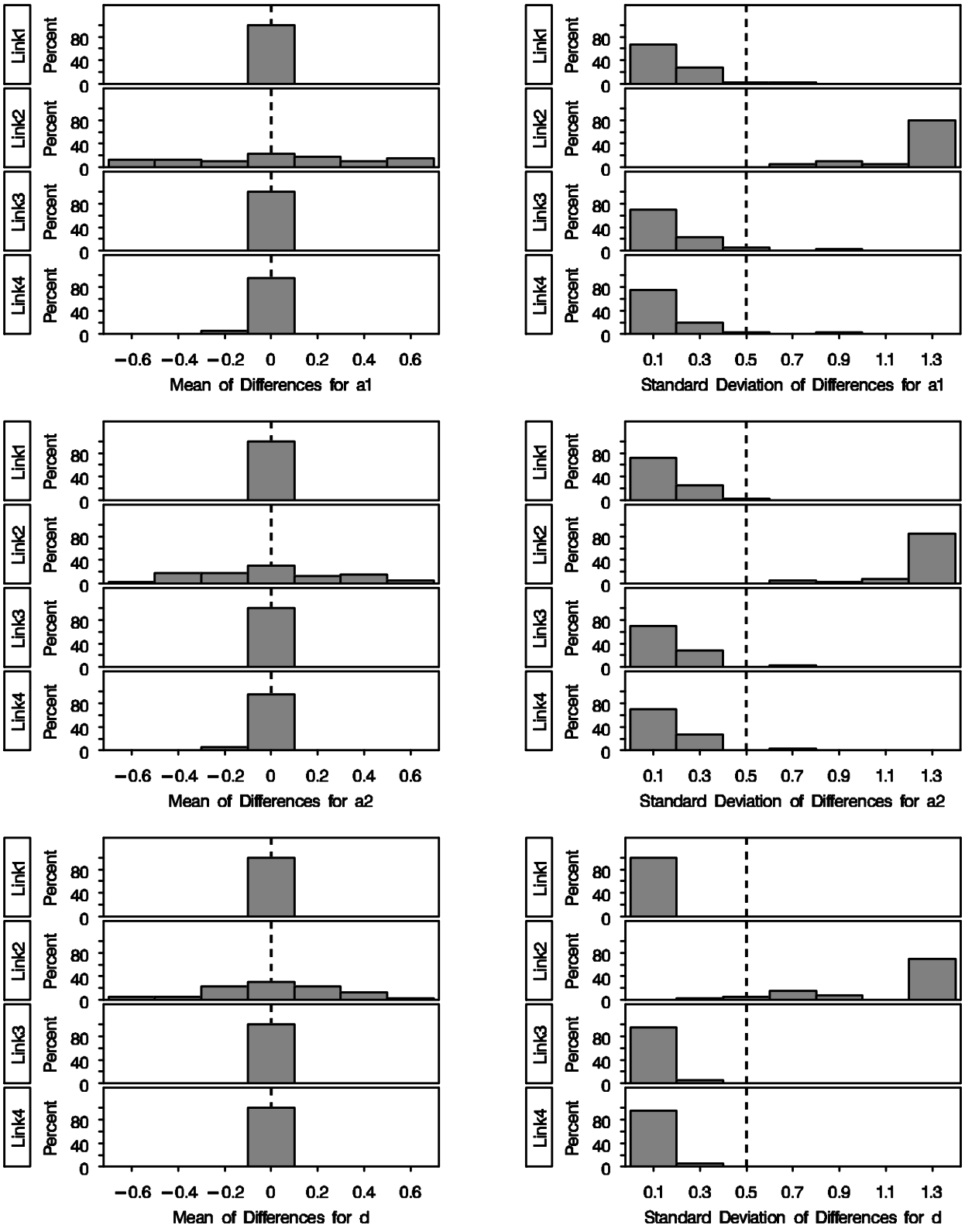


Figure A-41. Accuracy and stability for different linking methods (COM, n=40, N=1000, G1)

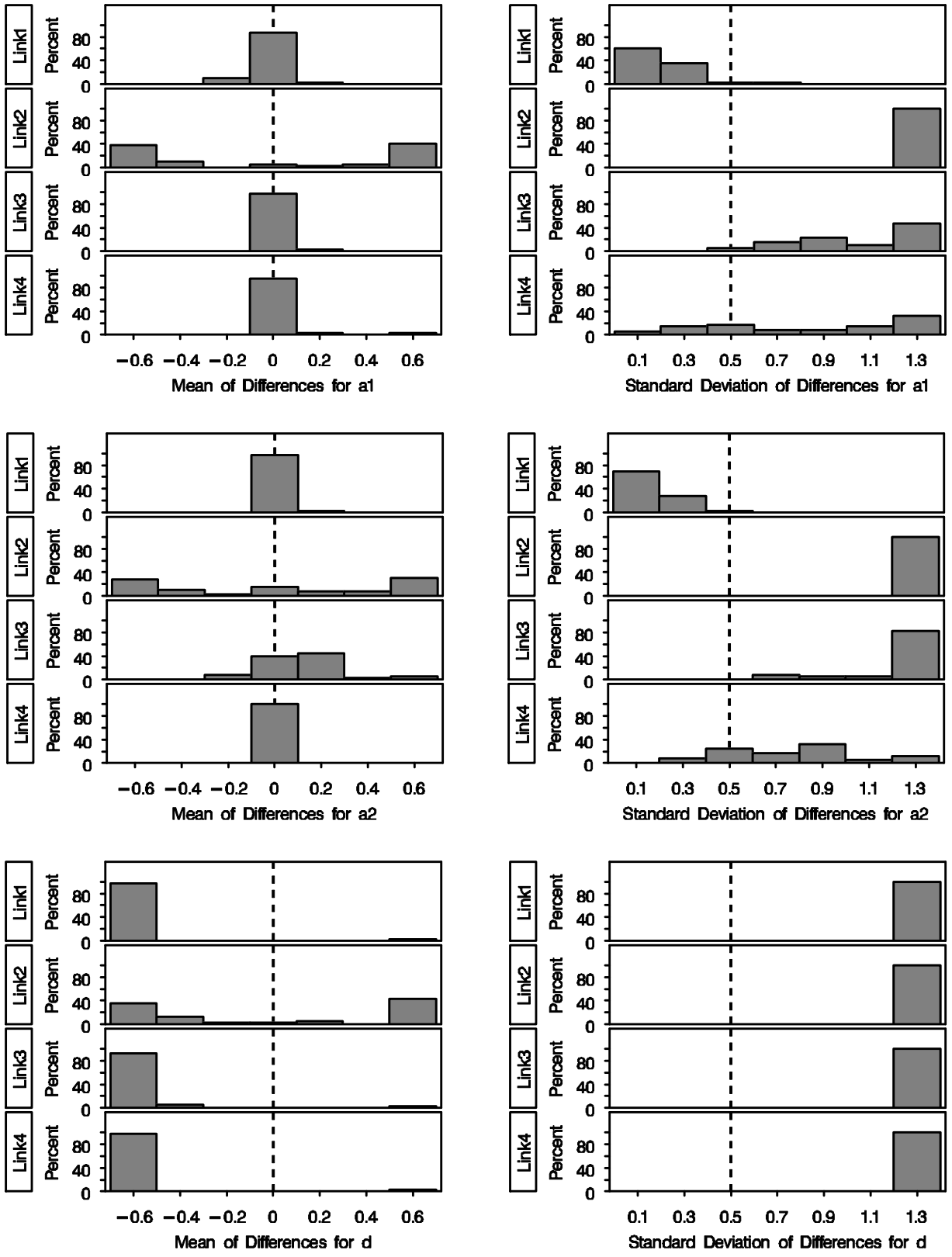


Figure A-42. Accuracy and stability for different linking methods (COM, n=40, N=1000, G2)

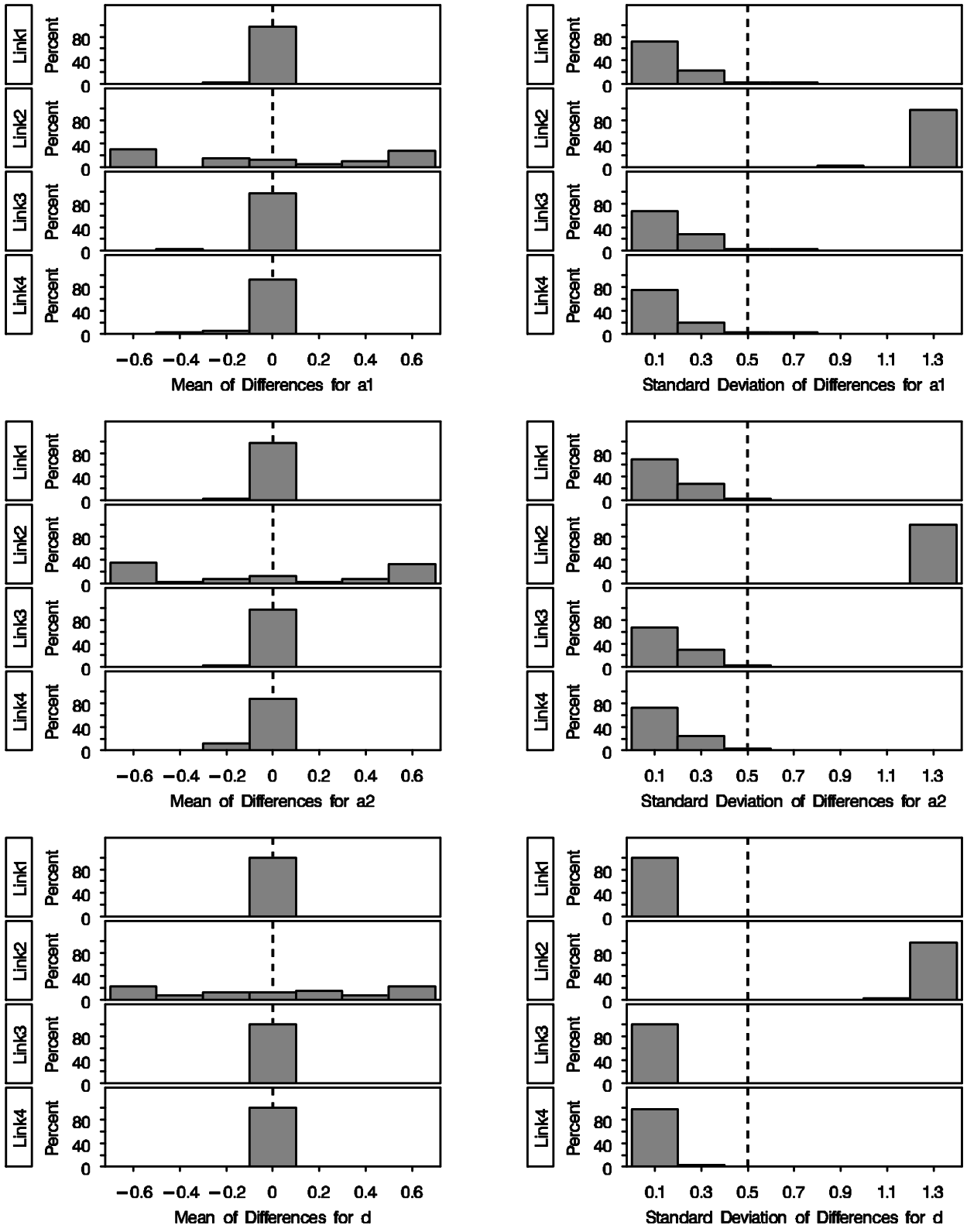


Figure A-43. Accuracy and stability for different linking methods (COM, n=40, N=1000, G3)

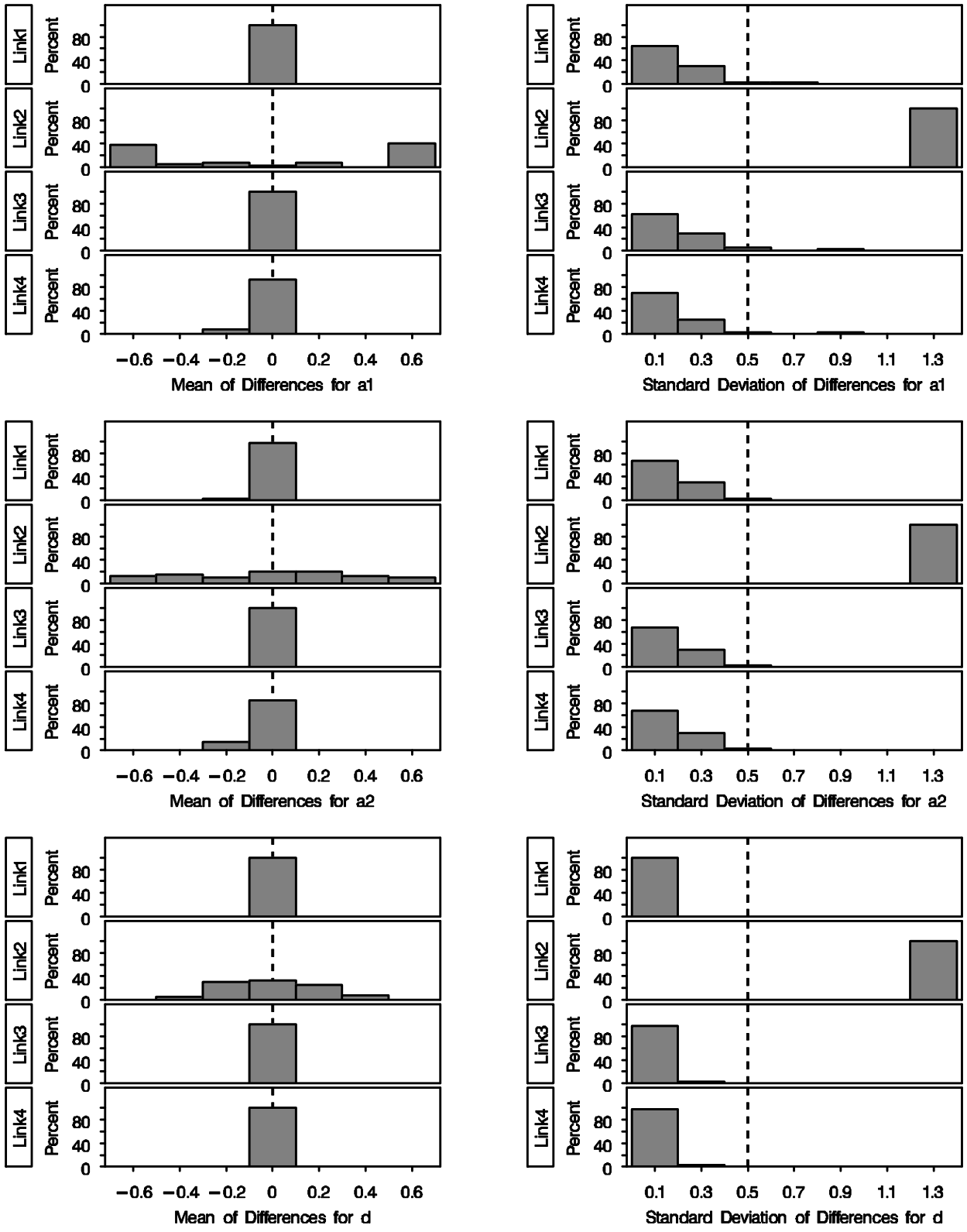


Figure A-44. Accuracy and stability for different linking methods (COM, n=40, N=1000, G4)

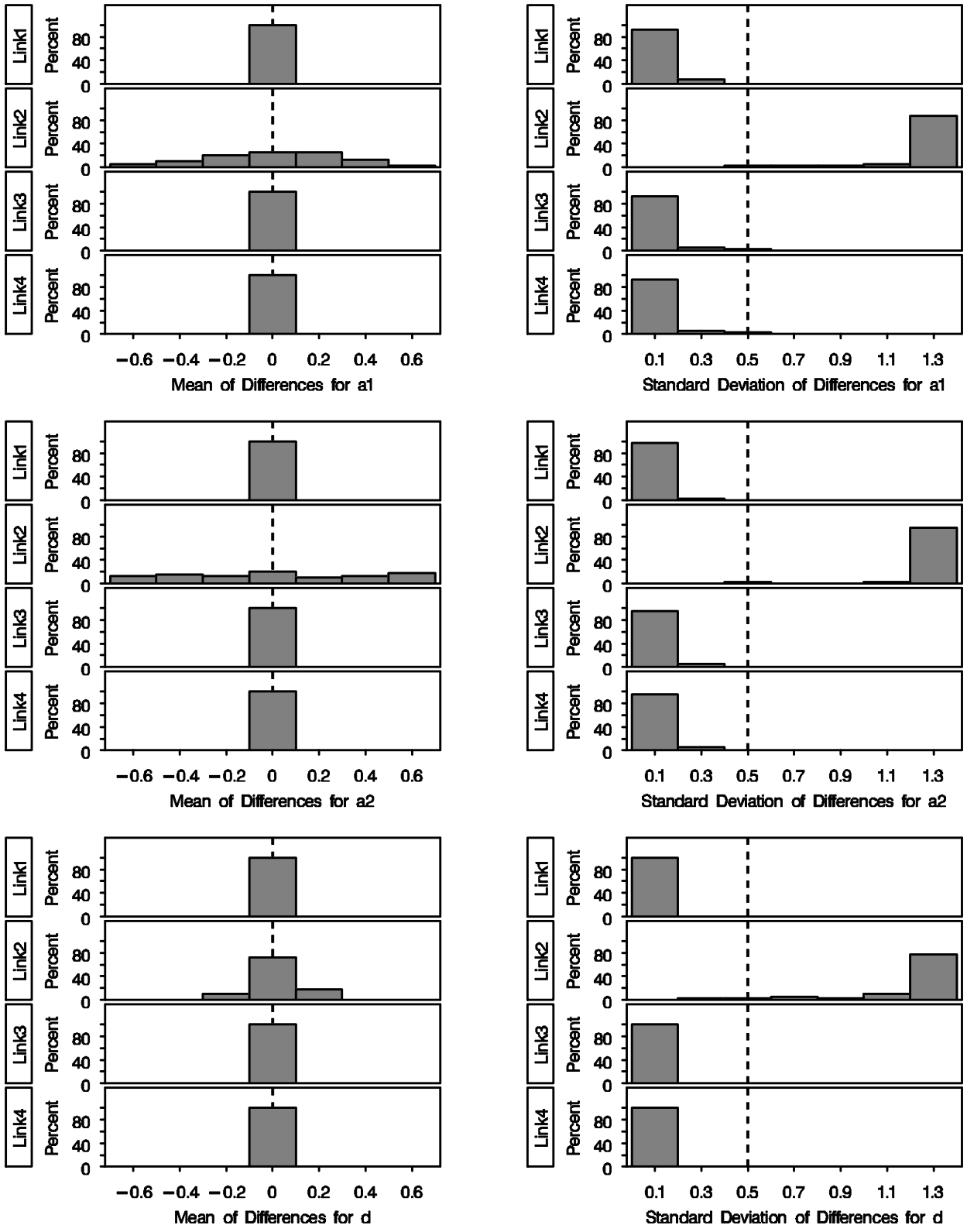


Figure A-45. Accuracy and stability for different linking methods (COM, n=40, N=2000, G1)

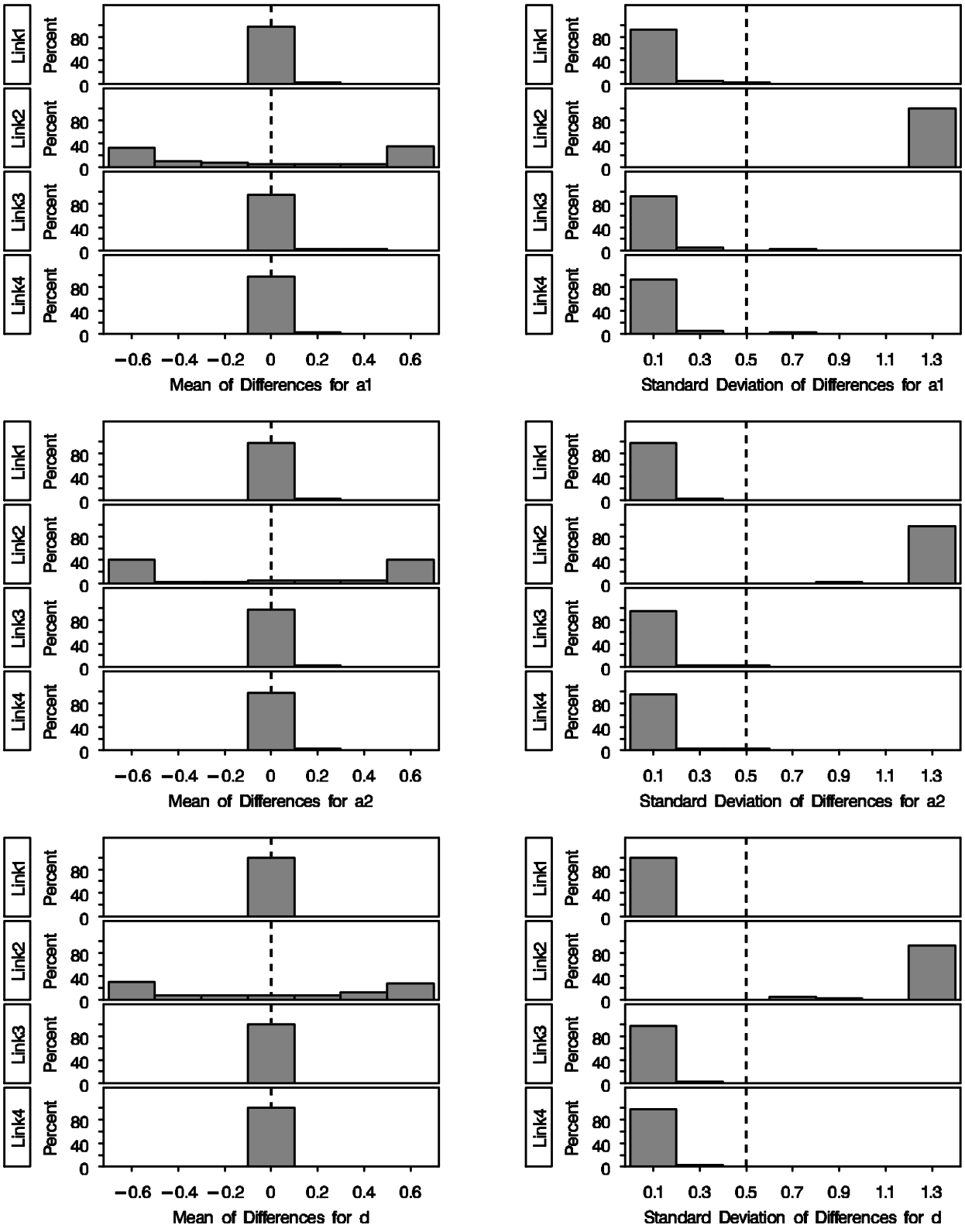


Figure A-46. Accuracy and stability for different linking methods (COM, n=40, N=2000, G2)

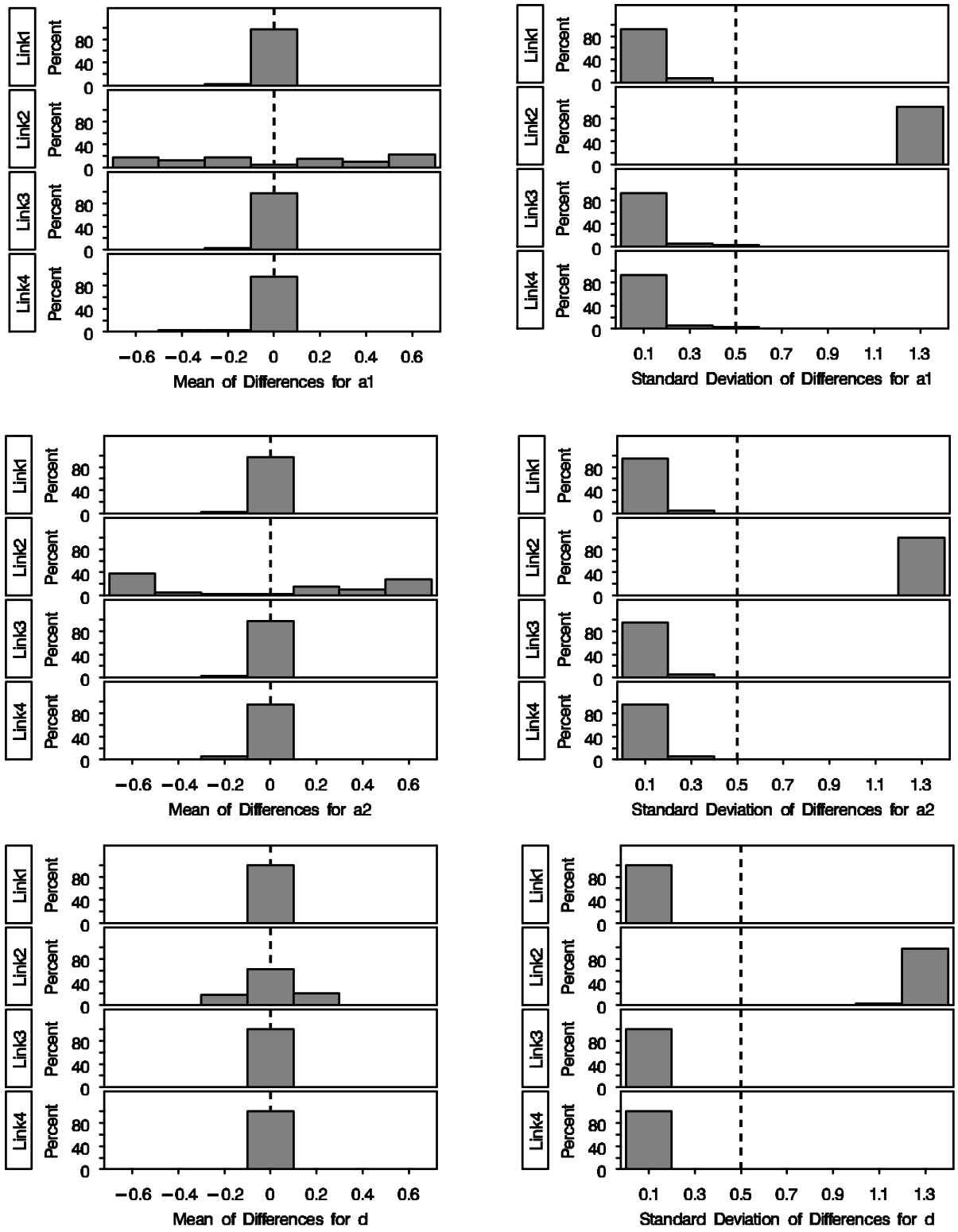


Figure A-47. Accuracy and stability for different linking methods (COM, n=40, N=2000, G3)

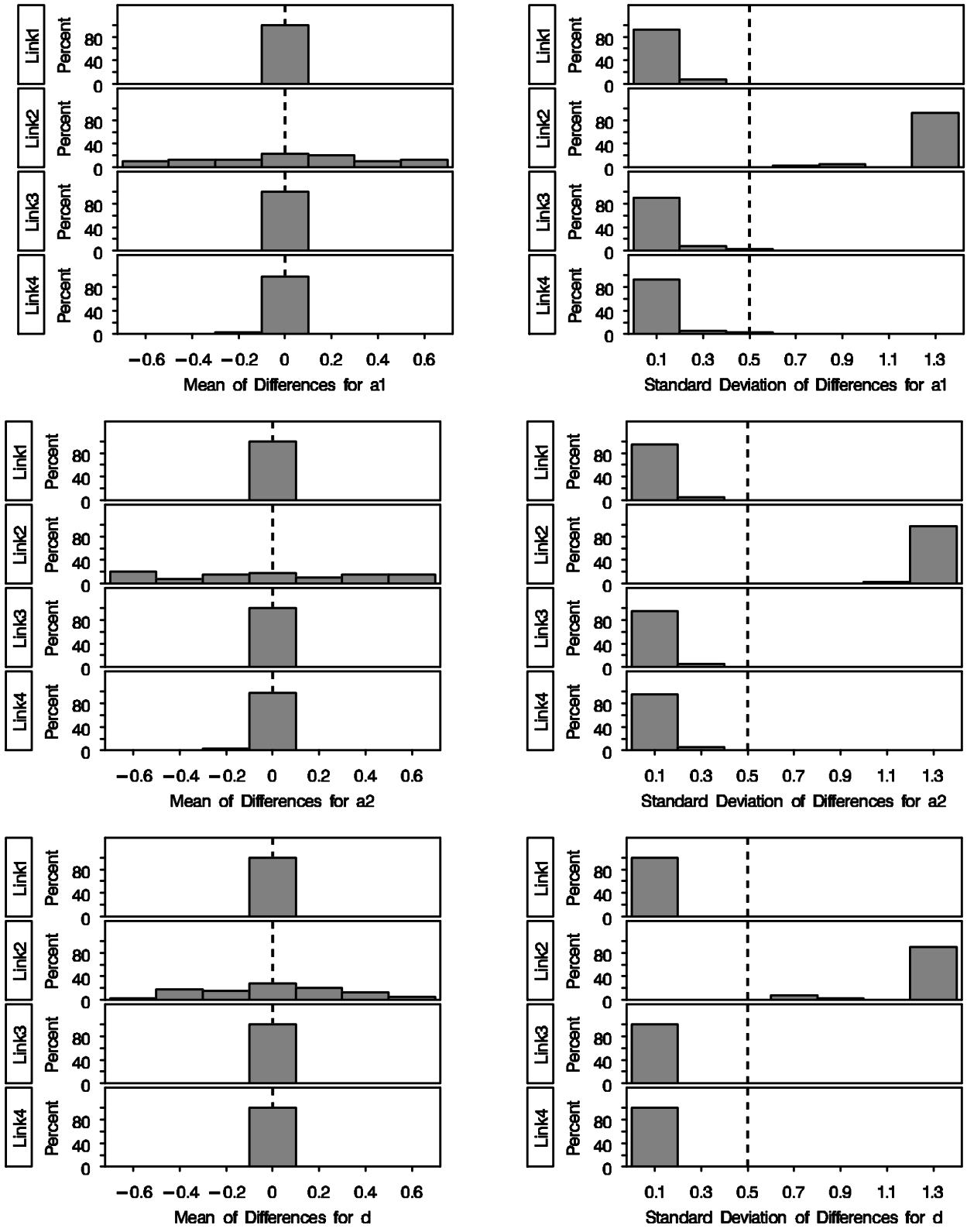


Figure A-48. Accuracy and stability for different linking methods (COM, n=40, N=2000, G4)

LIST OF REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255-278.

Ackerman, T. A. (1996). Graphic representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*, 311-329.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

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, 17*, 239-251.

Bateley, R. M., & Boss, M. W. (1993). The effects on parameter estimation of correlateddimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17*, 131-141.

Bedescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Journal of Educational Measurement, 22*, 13-20.

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

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (1999). *TESTFACT 3: Test scoring, items statistics, and full-information item factor analysis*. Chicago: Scientific Software International.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program.* Unpublished manuscript.

Cattell, R. B. (1978). *The scientific use of factor analysis*. New York: Plenum.

Cohen, A. S., & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*, 116-130.

Comery, A. L., & Lee, H. B. (1992). *A First course in factor analysis*. Hillsdale, NJ: Erlbaum.Cook, L. L., Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37-45.

Cook, L. L., Eignor, D. R., & Taft, H. (1981, April). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* (RR-85-38). Princeton, NJ: Educational Testing Service.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225-244.

Cureton, E. E., & D'Agostino, R. B. (1983). *Factor analysis: An applied approach*. Hillsdale, NJ: Erlbaum.

Davey, T. C., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20,* 405-416.

Divgi, D. R. (1985). A minimum chi-square methods for developing a common metric in item response theory. *Applied Psychological Measurement, 9*, 413-415.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education, 14*, 31-57.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis.*Multivariate Behavioral Research, 23,* 267-269.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gosz, J. K. & Walker, C. M. (2002, April). *An empirical comparison of simple versus complex multidimensional item response data.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Gosz, J. K., Walker, C. M. (2002, April). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM*. Paper presented at the Annual Meeting of the National Council for Measurement in Education (NCME), New Orleans, Louisiana.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw Hill.

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

Hason, B. A., & Beguin, 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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff Publishing.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26*, 337-349.

Hirsch, T. M. (1988). *Multidimensional equating*. Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25,* 39-52.

Kelderman, H. (1997). Loglinear multidimensional item response models for polytomously scored items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign. Urbana-Champaign, IL.

Kim, S. H., & Cohen, A. S. (1995). A minimum chi-square method for equating tests under the graded response model. *Applied Psychological Measurement, 19*, 167-176.

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*, 25-41.

Klein, L. W., & Kolen, M. J. (1985, April). *Effect of number of common items in common-item equating with nonrandom groups.* Paper presented at the annual meeting of American Educational Research Association, Chicago.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

Kolen, M. J. (2004a). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3-14.

Kolen, M. J. (2004b). Linking assessment: Concept and history. *Applied Psychological Measurement, 28,* 219-226.

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

Lautenschlager, G. J., Flaherty, V. L., & Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement, 54,* 21-31.

Lee, K., & Oshima, T. C. (1996). IPLINK: Multidimensional and unidimensional item parameter linking in item response theory. *Applied Psychological Measurement, 20*, 230.

Li, Y. H. (1997). An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric (Doctoral dissertation, University of Marryland, 1997). *Dissertation Abstract International,* UMI Number 9816494.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24,* 115-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

McDonald, R. P. (1999). *Test theory: a unified treatment.* Mahwah, NJ: Erlbaum.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report, ONR 83-2). Iowa City, IA: American College Testing Program.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary Psychometrics*. Mahwah, NJ: Lawrence Erlbaum Associates.

Min, K. S. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations.* Unpublished Dissertation, Michigan State University, East Lansing, MI.

Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education, 19*, 67-91.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1999). *POLYFACT version 2* [Computer program]. Princeton, NJ: Educational Testing Service.

Muraki, E, & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Muraki E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417-430.

Muthen, L. K., & Muthen, B. (1998). *MPLUS: The comprehensive modeling program for applied researcher: User's guide.* Los Angeles: Muthen & Muthen.

Oshima, T. C., Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237-248.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*, 357-373.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14,* 163-173.

Peterson, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Peterson, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (pp. 221-262). New York: American Council on Education and Macmillan.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (1991). The discriminating power of items that measure more than onedimension. *Applied Psychological Measurement, 15*, 361-373.

Reckase, M. D. (1997a). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

Rechase, M. D., & Martineau, J. (2004, October). *The vertical scaling of Science Achievement Tests.* Paper commissioned by the Committee on Test Design for K-12 Science Achievement, Center for Education, National Research Council.

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on liner equating of item characteristic curve parameters. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 135-146). New York: Academic.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1-30.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Schonemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika, 31*, 1-10.

Schonemann, P. H., & Carroll, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika, 35*, 245-255.

Simpson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Stone, C. A., & Yeh, C. C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams. *Educational and Psychological Measurement, 66*, 193-214.

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizon in testing* (pp. 13-30). New York: Academic.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

Thissen, D. (1991). *MULTILOG user's guide: multiple, categorical item analysis and testscoring using item response theory* [Computer program]. Chicago, IL: Scientific Software International.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 392-412.

Thompson, T. D., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT scale linking without common items or common examinees.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Vandenberg, R. J.,& Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practice, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70.

Wang, M. (1985). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT*. Unpublished manuscript.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (Research Report 87-24). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Eds.), *Educational Measurement (4th ed.).* West Port, CT: Praeger.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer program]. Chicago, IL: Scientific Software International.

BIOGRAPHICAL SKETCH

Youhua Wei was born in China. He received his B.Ed. in school education from Nanjing Normal University in 1992 and his M.Ed. in psychology from East China Normal University in 1995. From 1995 to 1997, he worked as a psychological counselor at Southeast University in Nanjing. From 1997 to 2001, he worked as a research associate at Shanghai Academy of Educational Sciences. In 2004, he earned his M.S. in research, measurement, and statistics from Texas A&M University in College Station. He began his doctoral study in research and evaluation methodology in the Department of Educational Psychology at the University of Florida in fall 2004. He was awarded the Ph.D. degree in August 2008.