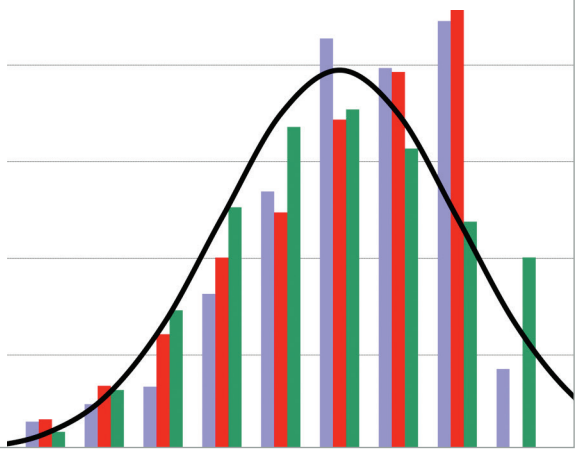
ACT Research
Report Series
2013 (2)



A Comparison of Four Linear Equating Methods for the Common-Item Nonequivalent Groups Design Using Simulation Methods

Anna Topczewski
Zhongmin Cui
David Woodruff
Hanwei Chen
Yu Fang

ACT®

For additional copies, write:

ACT Research Report Series
P.O. Box 168
Iowa City, IA 52243-0168

A Comparison of Four Linear Equating Methods for the Common-Item Nonequivalent Groups Design using Simulation Methods

Anna Topczewski
Zhongmin Cui
David Woodruff
Hanwei Chen
Yu Fang

Abstract

This paper investigates four methods of linear equating under the common item non-equivalent groups design. Three of the methods are well known: Tucker, Angoff-Levine, and Congeneric-Levine. A fourth method is presented as a variant of the Congeneric-Levine method. Using simulation data generated from the three-parameter logistic IRT model we compare the accuracy of the four methods under a variety of conditions involving group differences between the old and new groups. The sampling properties of the methods' parameter estimates are also investigated. The results indicate that the Tucker method is less accurate than the other three methods when group differences exist, especially when sample size is large (800). However, the Tucker method's gamma has the smallest sampling error, especially when sample size is small.

A Comparison of Four Linear Equating Methods for the Common-Item Nonequivalent Groups Design using Simulation Methods

Introduction

In high stakes testing situations a new form is administered on each test date with examinees possibly varying in achievement over test dates. This situation requires an equating method that can disentangle form differences (which form was easier/harder) from group differences (which group was higher/lower achieving). The common-item nonequivalent groups (CINEG) design can be used to separate out form and group differences through the use of a small common item set, also called an anchor test (Kolen & Brennan, 2004). Because the common items are given to both groups, the group difference on the common items is used to estimate the group difference between the two forms. There are two variations of the CINEG design; the *internal* design where the common item set contributes to the examinee's score and the *external* design where the common item set does not contribute to the examinee's score. Some authors refer to CINEG design (internal or external) as the non-equivalent groups with anchor test (NEAT) design (internal or external).

A difference between the CINEG design and the random groups equating design is that the former involves two groups possibly differing in achievement. The concept of a synthetic population as described by Braun and Holland (1982) and Angoff (1971, 1982) is used to combine the two groups into one group for estimating the equating relationship. The new group is given the weight w_i and the old group is given the weight w_0 where $w_i + w_0 = 1$. Usually both weights are given the value 0.5; alternatively, the weights $w_i = 1$ and $w_0 = 0$ can be used so that the synthetic group is defined as just the new group. The latter approach is used in this paper.

Several observed score linear equating methods for the CINEG design have been developed: the Tucker method (Gulliksen, 1950, pp. 299-301), the Angoff-Levine method

described by Angoff (1982, 1984) which incorporates Angoff's (1953) reliability formula, and the Congeneric-Levine method developed by Levine (1955) and further derived by Woodruff (1986). A new fourth method presented in this paper is a variation of the Congeneric-Levine method. The interested reader should also see Kane, Mroch, Suh, and Ripkey (2009a); Kane, Mroch, Suh, and Ripkey (2009b); Suh, Mroch, Kane, and Ripkey (2009); and Mroch, Suh, Kane, and Ripkey (2009), but note that what those papers refer to as the Levine observed score equating method is the Angoff-Levine method in this paper; they do not consider the Congeneric-Levine method. The results in those papers generally agree with the result found in this paper, which is that the Tucker method ignores measurement error and as a consequence is inaccurate when group differences exist.

Table 1 on the next page shows the layout and notation for the observed linear equating design used in this study. Note that *m* and μ denote sample mean and population mean, respectively, and *s* and σ are similarly defined for the standard deviation. A tilde denotes that an indirect estimate is needed due to missing data. The goal is to equate the new form *e* to the old form *o*, that is, to find a mapping of a score on the new form *e* to its comparable score on the old form *o*. This represents the external design. The internal design equates (*e* + *c*) to (*o* + *c*), where c is the anchor test of common items. This is done in the new group *i* using the indirect estimates $\tilde{m}(o_i)$ and $\tilde{s}(o_i)$ with the equation

$$\tilde{o}_i(e_i) = \left[\frac{\tilde{s}(o_i)}{s(e_i)}\right] e_i + \left[\tilde{m}(o_i) - \frac{\tilde{s}(o_i)}{s(e_i)} m(e_i)\right] \tag{1}$$

where when $w_i = 1$ and $w_0 = 0$

$$\tilde{m}(o_i) = m(o_0) + \gamma_{0,M}[m(c_i) - m(c_0)], \text{ and} \tag{2}$$

$$\tilde{s}(o_i) = \sqrt{s^2(o_0) + \gamma_{0,M}^2[s^2(c_i) - s^2(c_0)]}\,. \tag{3}$$

These indirect estimates depend on the coefficients $\gamma_{0,M}$, which differ by method. The first subscript 0 (zero) on γ denotes that it is computed using old group statistics and the second subscript *M* denotes the method.

Table 1.

The Common Item Non-Equivalent Group Design Used in This Paper

	Group	
Test	**Old Group 0** (Test *o* Administered)	**New Group *i*** (*i*=1, 2 ,3 ,4) Test *e* Administered
New Test *e* (Score based on unique even-numbered items)	**NO DATA** **Indirect Estimates Not Needed When** $w_0 = 0$	**DATA** **Direct Values:** Parameters: $[\mu(e_i), \sigma^2(e_i)]$ Statistics: $[m(e_i), s^2(e_i)]$
Old Test *o* (Score based on unique odd-numbered items)	**DATA** **Direct Values:** Parameters: $[\mu(o_0), \sigma^2(o_0)]$ Statistics: $[m(o_0), s^2(o_0)]$	**NO DATA** **Indirect Estimates Needed:** Parameters: $[\tilde{\mu}(o_i), \tilde{\sigma}^2(o_i)]$ Statistics: $[\tilde{m}(o_i), \tilde{s}^2(o_i)]$
Anchor Test *c* (Score on anchor test)	**DATA** **Direct Values:** Parameters: $[\mu(c_0), \sigma^2(c_0)]$ Statistics: $[m(c_0), s^2(c_0)]$	**DATA** **Direct Values:** Parameters: $[\mu(c_i), \sigma^2(c_i)]$ Statistics: $[m(c_i), s^2(c_i)]$

Let var denote variance, cov denote covariance, rel denote reliability, and $\alpha(o_0 \mid c_0)$ denote the slope of the linear regression of the unique items on the common items. The formulas for the gammas for each method are:

$$\gamma_{0,T} = \frac{\mathrm{cov}(o_0, c_0)}{\mathrm{var}(c_0)} = \alpha(o_0 \mid c_0), \tag{4}$$

$$\gamma_{0,AL} = \frac{\alpha(o_0 \mid c_0)}{\mathrm{rel}(c_0)} = \frac{\mathrm{rel}(o_0)}{\alpha(c_0 \mid o_0)} = \frac{\left[\mathrm{cov}(o_0, c_0) + \mathrm{var}(o_0)\right]}{\left[\mathrm{cov}(o_0, c_0) + \mathrm{var}(c_0)\right]}, \tag{5}$$

$$\gamma_{0,CL1} = \frac{\mathrm{cov}(o_0, c_0)}{\left[\mathrm{rel}(c_0)\,\mathrm{var}(c_0)\right]} = \frac{\alpha(o_0 \mid c_0)}{\mathrm{rel}(c_0)}, \text{ and} \tag{6}$$

$$\gamma_{0,CL2} = \frac{\left[\mathrm{rel}(o_0)\,\mathrm{var}(o_0)\right]}{\mathrm{cov}(o_0, c_0)} = \frac{\mathrm{rel}(o_0)}{\alpha(c_0 \mid o_0)} \tag{7}$$

where the method subscripts are as follows: T for Tucker, AL for Angoff-Levine, CL1 for the first Congeneric-Levine, and CL2 for the second Congeneric-Levine. In this paper reliability is estimated by coefficient alpha although other coefficients could be used. The formulas in equations (4) through (7) are for the external anchor situation, and they may be easily derived from the results given in Woodruff (1986). As Woodruff (1986) shows the gammas for the internal anchor situation can be found by adding unity to the external gamma estimates. For simplicity of notation and to illustrate their dependence on reliability, or lack thereof, all of the formulas given throughout are for the external anchor situation. However, the results of the simulation study presented below are for the more common internal anchor situation.

The rational for the development of the second Congeneric-Levine method is that the first Congeneric-Levine method depends on the reliability of the usually very short test composed of the common items. The second Congeneric-Levine method reverses the regression and depends on the reliability of the usually much longer test composed of the unique items. Reliability (coefficient alpha) estimates of long tests are usually more stable than reliability estimates of short tests because the distribution of coefficient alpha depends on both the number of items and the number of examinees (Feldt, Woodruff, & Salih, 1987). It can be shown under a congeneric

model (Woodruff, 1986) that equations (6) and (7) are just different estimates for the same parameter. Note that equations (5) and (6) are not equal. The right-most expression of equation (5) is only valid under the special assumption of a classically congeneric model (Woodruff, 1986).

The two Congeneric-Levine methods and the Angoff-Levine method explicitly depend on a reliability coefficient, although under the assumptions of the classically congeneric model (Kolen & Brennan, 2004; Woodruff(1986)), upon which the Angoff-Levine method depends, an estimate of the reliability of the common items (or unique items) can be found without resort to item scores (Woodruff, 1986). The only method that does not consider reliability is the Tucker method. In particular, the Tucker method uses a simple observed score linear regression slope parameter, whereas the other three methods use a slope parameter that is corrected for measurement error in the predictor (Buonaccorsi, 2010, p85). As has been stated, when group differences exist in means or variances, the Tucker method give less weight to those differences, and so produces less accurate equating results.

Establishing a Criterion for Comparison

In the present study, a simulation method creates data for all six of the data cells in Table 1 thereby enabling a comparison between indirect estimates and their corresponding direct values. The total error for the mean and variance of the indirect estimates can be defined as the sum of two component errors: the first is the error due to the method and the second is the error due to sampling. They are defined as follows when a sample is collected from a population:

$$\tilde{m}(o_i) - \mu(o_i) = \left[\tilde{m}(o_i) - m(o_i)\right] + \left[m(o_i) - \mu(o_i)\right] \text{ and} \tag{8}$$

$$\tilde{s}^2(o_i) - \sigma^2(o_i) = \left[\tilde{s}^2(o_i) - s^2(o_i)\right] + \left[s^2(o_i) - \sigma^2(o_i)\right]. \tag{9}$$

When population data is available to compute the indirect estimates, error due to sampling is eliminated and error due to method equals the total error for the mean and variance for the indirect estimates. They are defined as follows:

$$\text{method\_err}(\tilde{\mu}) = \tilde{\mu}(o_i) - \mu(o_i) \text{ and} \tag{10}$$

$$\text{method\_err}(\tilde{\sigma}^2) = \tilde{\sigma}^2(o_i) - \sigma^2(o_i). \tag{11}$$

The methods differ in their estimates for gamma which is the coefficient that steps-up the group differences on the common items to the scale of the total test. To quantify this accuracy, the errors due to method, that is, the first term on the right sides of equations (8) and (9) and the formulas in equations (10) and (11) will be evaluated and compared across the four equating methods. However these raw values are not ideal for direct comparison for two reasons: first method error can be positive or negative and second method error is metric dependent. To remedy these two problems the absolute value of the method error can be taken, and then the relative error of this quantity can be found producing an absolute relative error (A-RLTER) for the method. For both the mean and variance indirect estimates A-RLTER due to the method are defined as follows for the population and sample data respectively:

$$\text{A-RLTER}(\tilde{\mu}) = \frac{abs\left[\tilde{\mu}(o_i) - \mu(o_i)\right]}{\mu(o_i)} * 100, \tag{12}$$

$$\text{A-RLTER}(\tilde{\sigma}^2) = \frac{abs\left[\tilde{\sigma}^2(o_i) - \sigma^2(o_i)\right]}{\sigma^2(o_i)} * 100, \tag{13}$$

$$\text{A-RLTER}(\tilde{m}) = \frac{abs\left[\tilde{m}(o_i) - m(o_i)\right]}{m(o_i)} * 100, \text{ and} \tag{14}$$

$$\text{A-RLTER}(\tilde{s}^2) = \frac{abs\left[\tilde{s}^2(o_i) - s^2(o_i)\right]}{s^2(o_i)} * 100. \tag{15}$$

The present study is interested in assessing how accurately each of the four equating methods, Tucker, Angoff-Levine, Congeneric-Levine 1 and Congeneric-Levine 2 estimate the indirect estimates. For population data mean and variance A-RLTERs, equations (12) and (13) will be computed and compared across the four equating methods. For sample data, the average of the mean and variance A-RLTERs, equations (14) and (15), will be taken across replications and then compared across the four equating methods.

Data Generation and Methodology

The simulation data used in this study were based on the item parameters of 75 items from a nationally published English test and 60 items from a nationally published Mathematics test. All items were multiple choice items. The three-parameter logistic (3PL) item response (IRT) model and a random sample of 10,000 examinees were used to estimate the item parameters for the two tests. The BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003) computer program was used for the estimation. The mean and standard deviation (in parentheses) of the estimated IRT parameters for each test are shown in Table A-1 in Appendix A.

Simulated Data

Population data. A random number generator was used to generate normally distributed theta (ability) values under four conditions: English with a small group differences (E1), English with large group differences (E2), Mathematics with small group differences (M1), and Mathematics with large group differences (M2). Within a condition, a population dataset of 400,000 examinee theta values (referred to as Old Group) and four additional population datasets of 400,000 examinee theta values each (referred to as New Groups 1 to 4 respectively) were created based on the manipulation of the new group in the following manner: no change in mean

and standard deviation, increase in mean and no change in standard deviation, no change in mean and an increase in standard deviation, and increase in mean and standard deviation. Therefore 5 datasets were created within each condition for a total of 20 total simulated examinee theta value datasets. See Table A-2 in Appendix A for the exact values of the means and standard deviations for the 20 conditions. Using the estimated item parameters and generated theta values for each of these 20 datasets, examinee dichotomously scored item responses were generated for each item using the 3PL model. In total, 8,000,000 examinee observations were generated.

Sample data. Random samples were drawn from each of the 20 simulated population datasets containing the examinee's dichotomously scored item responses. Sample sizes of 200, 400 and 800 were drawn 2000 times. The random sampling was done with replacement.

Form Creation

For each of the English datasets (population and sample) two forms each with a total of 43 items (unique plus common) were created by using odd numbered items as the old form and even number items as the new form with the exception of item 75 which was included on both forms to obtain equal length forms. To create a common item set, five additional items from each form were selected and crossed with the other form, for a total of 11 common items. See Table A-3 in Appendix A for more details.

Similarly for each of the Mathematics datasets (population and sample) old and new forms each with a total of 35 items (unique plus common) were created by using odd numbered items as the old form and even numbered items as the new form. To create a common item set, five items from each form were selected and crossed with the other form, for a total of 10 common items between the forms. See Table A-3 in Appendix A for more details.

Short tests and small sample sizes were used for two reasons. The first is such situations can be found in practice, and the second is that such situations stress the methodology.

Equating Relationship/Indirect Estimates

The indirect estimates, equations (2) and (3) of the equating relationship, where a new group who took the even form was equated to the old group who took the odd form, were found for all population and sample datasets using the four previously described equating methods. Note again that although all equations are for the external anchor situation, except where noted the results presented in the next section are for the more common internal anchor situation where o represents $o + c$ and e represents $e + c$.

Results

Population Data

Parameter values. Mean, variance, covariance, and reliability (coefficient alpha) parameters were calculated for each group (Old Group and New Groups 1 to 4) within each condition (E1, M1, E2, and M2) and can be found in Tables A-4 through A-7 in Appendix A. In addition and following Woodruff (1986) two group versions of the congeneric model underlying Levine-congeneric equating and the classically congeneric model underlying Angoff-Levine equating were fit to each of the 20 groups by condition combinations using the Mplus program (Muthen & Muthen, 2010). Both models fit the data fairly well, but the congeneric model fit significantly better than the classically congeneric model which is more restrictive than the congeneric model.

In Table A-8 in Appendix A, mean observed differences between the forms and groups can be found and were calculated as Observed difference = Form difference for new group + Group difference on old form, i.e.,

$$[\mu(e_i) - \mu(o_0)] = [\mu(e_i) - \mu(o_i)] + [\mu(o_i) - \mu(o_0)]. \quad (16)$$

The mean difference of the groups on the common items is also given. Variance differences are also given in Table A-8 in Appendix A and were calculated in a manner similar to the mean differences calculation. The results show the mean and variance group differences are consistent with what would be expected given the manipulation of the data shown in Table A-2 in Appendix A.

Comparison of methods. Figures B-1 and B-2 in Appendix B display the A-RLTER of the mean and variance indirect estimates for each of the four equating methods. As can be seen when there are only minimal group differences in the mean and variance (New Group 1) all methods have very low A-RLTER for both the mean and the variance indirect estimates. When only mean group differences are observed (New Group 2) the Congeneric-Levine 2 and Congeneric-Levine 1 methods have the lowest A-RLTER for all mean indirect estimates for small and large group differences respectively, and Tucker has the lowest A-RLTER for the variance indirect estimates. When only variance group differences are observed (New Group 3) the Tucker methods has the lowest A-RLTER for all mean indirect estimates, the Congeneric-Levine 1 method has the lowest A-RLTER for the English indirect estimates of variance, and the Angoff-Levine method has the lowest A-RLTER for the Mathematics indirect estimates of variance. When both mean and variance group differences (New Group 4) are observed the Congeneric-Levine 1 method has the lowest A-RLTER for the English mean and variance indirect estimates. Although the Congeneric-Levine 2 methods has the lowest A-RLTER for the Mathematics mean indirect estimates, the Angoff-Levine method has the lowest A-RLTER for the Mathematics variance indirect estimates.

Sample Data

Descriptive statistics. Mean and variance observed differences between the forms and groups were calculated for the sample data as previously described for the population data and were found to be comparable with their population counterparts. These results are presented Table A-9 in Appendix A. Mean, variance, covariance, and reliability (coefficient alpha) statistics for each group (Old Group and New Groups 1 to 4) within each condition (E1, M1, E2, and M2) and for each replication within each sample size (N=200, 400, 800) are not presented because of their length

Comparison of methods. The A-RLTER of the mean and variance indirect estimates of each of the four equating methods for each sample size replication was calculated and the average of these statistics was then taken across the 2000 replications. The Tucker method had the lowest mean A-RLTER for the mean indirect estimates when there was only minimal mean group differences (New Group 1 and 3) and had the lowest mean A-RLTER for the variance indirect estimates when there was minimal variance group differences (New Group 1 and 2).

Figure B-3 in Appendix B shows the mean A-RLTER of the sample mean indirect estimates when group means differ (New Group 2 and 4). When N=200 and there are small mean group differences (E1 and M1) the Tucker or Angoff-Levine methods have the lowest mean A-RLTER, but when N=400 or 800 the Angoff-Levine or Congeneric-Levine 2 methods have the lowest mean A-RTLER. When there are larger mean group differences (E2 and M2) the Angoff-Levine or Congeneric-Levine 2 methods have the lowest mean A-RTLER for all sample sizes, and the Tucker method has much larger mean A-RLTER than all other methods, and its A-RLTER does not always decrease as sample size increases, because it is biased.

Figure B-4 in Appendix B shows the mean A-RLTER of the sample variance indirect estimates when group variances differ (New Group 3 and 4). When N=200 or 400 and there are small English variance group differences (E1) the Tucker method has the smallest mean A-RLTER and when N=800 the Angoff-Levine has the lowest mean A-RLTER. When N=200 and there are Mathematics small variance group differences (M1) the Tucker method has the smallest mean A-RLTER and when N=400 or 800 the Angoff-Levine or Congeneric-Levine 2 method has the lowest mean A-RLTER. For all sample sizes when there are larger group differences (E2 and M2), the Angoff-Levine or Congeneric-Levine 2 methods have the lowest mean A-RTLER, and the Tucker method has much larger mean A-RLTER than all other methods. Moreover, the A-RLTER for the Tucker method does not always decrease as sample size increases, because it is biased.

Sample gamma coefficients. The sampling properties of the gamma coefficients were investigated. Let $g_{0,M}$ denote the sample gamma coefficient where the subscript 0 (zero) denotes that it was computed in the old group and the subscript *M* indicates the method. Let $\gamma_{0,M}$ denote its parametric counterpart. In this study mean squared error (MSE) for the sample gammas is defined as

$$MSE(g_{0,M}) = (2000)^{-1} \sum_{i=1}^{2000} (g_{0,M} - \gamma_{0,M})^2 . \qquad (17)$$

The MSE can be decomposed into two additive parts, sampling error and bias, as defined by the following equation

$$MSE(g_{0,M}) = VAR(g_{0,M}) + (\bar{g}_{0,M} - \gamma_{0,M})^2 \qquad (18)$$

where $\bar{g}_{0,M}$ is the sample mean of the gamma coefficient and $(\bar{g}_{0,M} - \gamma_{0,M})^2$ is squared bias. All three of these statistics were calculated for all groups, all conditions, and all sample sizes and the results can be found in Figures B-5 through B-7 in Appendix B.

First, it should be emphasized that for the sample sizes considered in this study, the MSE for all four methods' gammas are extremely small although the MSE for the Congeneric-Levine 1 method is relatively larger than the MSEs for the other three methods. The population gamma values have an approximate range of 2.5 to 4.1 across all 20 groups by conditions combinations, and all MSEs are less than 0.1. The relatively larger MSEs for the Congeneric-Levine 1 method is likely due to the denominator of the Congeneric-Levine 1 method involving the reliability of the relatively short anchor test because the distribution of coefficient alpha depends both on the sample size and the number of items (Feldt, Woodruff, & Salih, 1987). The relationships among the MSEs also hold for the random sampling error and bias with all bias values being less than .0012.

Figure B-5 in Appendix B shows that the Tucker method's sample gamma coefficients have the lowest random error followed by the Angoff-Levine, Congeneric-Levine 2, and Congeneric-Levine 1. Figure B-6 in Appendix B shows that when there are small group differences (E1 and M1) and when N=200 the Tucker method's sample gamma coefficients have the smallest squared bias. Otherwise all methods' sample gamma coefficients have relatively small squared bias, although the Congeneric-Levine 1 method's sample gamma coefficients squared bias is relatively larger than the other methods squared bias especially at the smallest sample size (N=200). The non-Tucker gamma coefficients usually have larger squared bias at N=200 but then tend to get smaller and closer to the squared bias for the Tucker method gamma at the 400 and 800 sample sizes. Figure B-7 in Appendix B shows that the Tuckers method's

sample gamma coefficients have the lowest MSE followed by the Angoff-Levine, Congeneric-Levine 2, and Congeneric-Levine 1, respectfully.

Discussion

This study assessed, under a variety of conditions, how accurately each of the four linear equating methods estimated the indirect means and variances needed in computing the equating relationship. As can be seen from equations (2) and (3) when there are no group differences in either mean or variance the choice of method is of little consequence. However, when group differences exist in means, variances, or both means and variances the choice of method does make a difference.

When sample size is small (N=200) and group differences are modest the Tucker method is either comparable to or slightly more accurate than the other methods, but as sample size increases (N=800) the Tucker method becomes less accurate than the other three methods. When group differences are moderate the Tucker method is less accurate than the other methods and this inaccuracy increases as sample size increases. The reason for this behavior is that the gamma coefficient in the Tucker method equals the linear regression slope of the unique items test score on the anchor items test ignoring measurement error in the anchor test score (Woodruff, 1986). The other three models use a regression model where the regression slope is corrected (disattenuated) for measurement error (Woodruff, 1986; Buonaccorsi, 2010). The value of the Tucker gamma will always be less than the gamma values of the other three methods and so will underestimate group differences when they exist.

When group differences exist and the Tucker method is contraindicated, one of the three other methods should be used. As Woodruff (1989) shows, the Angoff-Levine method depends crucially on the assumption of unity for the disattenuated correlation between the common item

test and the non-common item test, and also the assumption of classical parallelism. If there is any doubt regarding either assumption, then one of the two Congeneric-Levine methods is recommended. If the anchor test is short, with fewer than 20 items, then the second Congeneric-Levine method is recommended because it requires an estimate of the common item test reliability. If the anchor test is longer, with 20 or more items, then the first Congeneric-Levine method, which requires an estimate of the anchor test reliability, can be used. All reliability coefficients are lower bounds for reliability, but as test length increases the gap usually decreases (Woodruff & Wu, 2012). However, the second Congeneric-Levine method has better sampling properties as is evidenced in Figures B-5 through B-7.

Coefficient alpha is used as the reliability estimate in this paper. However, the generated data do not conform to an essentially tau equivalent model and so coefficient alpha is a lower bound to reliability. The simulations show that it is better to use an approximate reliability coefficient to correct the linear regression slope (non-Tucker methods) than to make no correction at all (Tucker method) when group differences exist and sample size is moderately large (800). At the larger sample size the Tucker method is rarely substantially more accurate than the other methods when group differences are small and at times much worse when group differences are large.

The results of this study can assist researchers in choosing the most accurate linear equating method for the often used common item nonequivalent groups design. Research in this area helps ensure the comparability of new test forms to old test forms thereby insuring fairness in testing over time and multiple forms.

References

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18,* 1-14.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P.W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55-70). New York: Academic Press.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton NJ: Educational Testing Service. [Reprint of chapter in R. L. Thorndike (Eds.), *Educational measurement* (2nd ed.). Washington DC: American Council on Education, 1971.]

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS procedures. In P.W. Holland & D.B. Rubin (Eds.), *Test equating* (pp. 9-50). New York: Academic Press.

Buonaccorsi, J. P. (2010). *Measurement Error*: *Models, methods, and applications.* Boca Raton: Chapman & Hall/CRC.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, 11, 93–103.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley.

Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009a). Potential bias in linear equating due to regression artifacts. *Measurement*, *7*, 123-124.

Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009b). Linear equating for the NEAT design: parameter substitution models and chained linear relationship models. *Measurement*, *7*, 125-146.

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

Levine, R. (1955). *Equating the score scales of alternative forms administered to samples of different ability* (ETS Research Bulletin No. 55-23). Princeton, NJ: Educational Testing Service.

Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement*, *7*, 174-193.

Muthen, L. K., & Muthen, B. O. (2010). *Mplus VERSION 6* [computer program]. Los Angeles: Muthen & Muthen.

Suh, Y., Mroch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An empirical comparison of five linear equating methods. *Measurement*, *7*, 147-173.

Woodruff, D. J. (1986). Derivations of observed score linear equating methods based on test score models for the common-item nonequivalent-populations design. *Journal of Educational Statistics, 11,* 245-257.

Woodruff, D. J. (1989). A comparison of three linear equating methods for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 13,* 257-261.

Woodruff, D. J., & Wu, Y, F. (2012). *Statistical considerations in choosing a test reliability coefficient*. (ACT Research Report 2012 (10)). Iowa City, IA: ACT.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago: Scientific Software.

Appendix A

Tables A-1 – A-8

Table A-1.

Mean and Standard Deviation of IRT Parameter Estimates

Test	$\hat{a}$	$\hat{b}$	$\hat{c}$
English	0.927 (0.303)	0.067 (0.756)	0.210 (0.094)
Mathematics	1.108 (0.380)	0.014 (1.048)	0.171 (0.083)

Table A-2.

Distributions of Examinee Theta Values

Conditions	Old Group	New Group 1	New Group 2	New Group 3	New Group 4
English small group differences (E1)	N(0, 1)	N(0, 1)	N(0.1, 1)	N(0, 1.1)	N(0.1, 1.1)
Mathematics small group differences (M1)	N(0, 1)	N(0, 1)	N(0.1, 1)	N(0, 1.1)	N(0.1, 1.1)
English large group differences (E2)	N(0, 1)	N(0, 1)	N(0.25, 1)	N(0, 1.25)	N(0.25, 1.25)
Mathematics large group differences (M2)	N(0, 1)	N(0, 1)	N(0.25, 1)	N(0, 1.25)	N(0.25, 1.25)

The given normal distribution variability parameter values are standard deviations not variances.

Table A-3.

Items on Each Form

Form	Unique Items	Number of Unique Items	Common Items	Number of Common Items
English Odd	3, 5, 7, 9, 11, 13, 17, 19, 21, 23, 25, 27, 31, 33, 35, 37, 39, 41, 45, 47, 49, 51, 53, 55, 59, 61, 63, 65, 67, 69, 71, 73	32	1, 2, 15, 16, 29, 30, 43, 44, 57, 58, 75	11
English Even	4, 6, 8, 10, 12, 14, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 46, 48, 50, 52, 54, 56, 60, 62, 64, 66, 68, 70, 72, 74	32	1, 2, 15, 16, 29, 30, 43, 44, 57, 58, 75	11
Mathematics Odd	3, 5, 7, 9, 11, 15, 17, 19, 21, 23, 27, 29, 31, 33, 35, 39, 41, 43, 45, 47, 51, 53, 55, 57, 59	25	1, 12, 13, 24, 25, 36, 37, 48, 49, 60	10
Mathematics Even	2, 4, 6, 8, 10, 14, 16, 18, 20, 22, 26, 28, 30, 32, 34, 38, 40, 42, 44, 46, 50, 52, 54, 56, 58, 60	25	1, 12, 13, 24, 25, 36, 37, 48, 49, 60	10

Table A-4.

Form and Group Population Parameters for E1

Parameters	Old Group	New Group 1	New Group 2	New Group 3	New Group 4
Mean					
New Form	25.0067	25.0112	25.7733	25.0622	25.8411
Old Form	25.6837	25.6935	26.4681	25.7175	26.4956
Common Items	6.9401	6.9425	7.1320	6.9331	7.1273
Variance					
New Form	69.0862	68.9020	69.8301	77.0800	77.6947
Old Form	71.6085	71.4079	71.7994	79.4970	79.5693
Common Items	5.8604	5.8652	5.8085	6.3262	6.2636
Covariance*					
New, Common	11.3690	11.3454	11.4905	12.8838	12.9555
Old, Common	11.7444	11.7194	11.7811	13.2434	13.2593
Reliability- Alpha*					
New Form	0.8502	0.8495	0.8547	0.8674	0.8712
Old Form	0.8600	0.8594	0.8633	0.8754	0.8781
Common Items	0.6398	0.6404	0.6450	0.6720	0.6762

*These are Parameters for the External Design

Table A-5.

Form and Group Population Parameters for M1

Parameters	Old Group	New Group 1	New Group 2	New Group 3	New Group 4
Mean					
New Form	20.1677	20.1719	20.8421	20.1998	20.8421
Old Form	19.4821	19.4854	20.1357	19.5482	20.1662
Common Items	5.2349	5.2358	5.4398	5.2671	5.4590
Variance					
New Form	49.1057	49.2475	49.4046	55.1614	55.2550
Old Form	46.3135	46.5419	47.0368	52.3904	52.8914
Common Items	5.5516	5.5827	5.6466	6.0817	6.1539
Covariance*					
New, Common	9.1422	9.1875	9.2906	10.4304	10.5258
Old, Common	8.7337	8.7982	8.9520	10.0309	10.1929
Reliability- Alpha*					
New Form	0.8333	0.8333	0.8361	0.8524	0.8546
Old Form	0.8166	0.8173	0.8219	0.8389	0.8426
Common Items	0.7044	0.7061	0.7136	0.7335	0.7404

*These are Parameters for the External Design

Table A-6.

Form and Group Population Parameters for E2

Parameters	Old Group	New Group 1	New Group 2	New Group 3	New Group 4
Mean					
New Form	24.9963	25.0211	26.9461	25.1444	26.9274
Old Form	25.6768	25.6982	27.6656	25.7423	27.5343
Common Items	6.9354	6.9405	7.4184	6.9240	7.3606
Variance					
New Form	68.8447	68.9717	69.8153	88.5608	89.6401
Old Form	71.4930	71.5234	70.9721	90.6783	90.5503
Common Items	5.8673	5.8689	5.6410	7.0004	6.7833
Covariance*					
New, Common	11.3448	11.3691	11.4135	15.0678	15.1520
Old, Common	11.7196	11.7334	11.6119	15.3935	15.3200
Reliability- Alpha*					
New Form	0.8493	0.8497	0.8598	0.8868	0.8943
Old Form	0.8598	0.8599	0.8669	0.8927	0.8986
Common Items	0.6401	0.6404	0.6484	0.7109	0.7181

*These are Parameters for the External Design

Table A-7.

Form and Group Population Parameters for M2

Parameters	Old Group	New Group 1	New Group 2	New Group 3	New Group 4
Mean					
New Form	20.1815	20.1725	21.8224	20.2520	21.7588
Old Form	19.4959	19.4812	21.0908	19.6389	21.1157
Common Items	5.2410	5.2353	5.7389	5.3115	5.7684
Variance					
New Form	49.1904	49.0154	49.2462	64.0966	64.1083
Old Form	46.4522	46.4532	47.4668	61.1719	62.2255
Common Items	5.5750	5.5567	5.7294	6.8340	6.9800
Covariance*					
New, Common	9.1747	9.1294	9.3680	12.3168	12.5077
Old, Common	8.7718	8.7634	9.1080	11.9029	12.2452
Reliability- Alpha*					
New Form	0.8331	0.8327	0.8392	0.8750	0.8795
Old Form	0.8171	0.8175	0.8278	0.8637	0.8711
Common Items	0.7054	0.7048	0.7242	0.7681	0.7817

*These are Parameters for the External Design

Table A-8.

Observed, Form, Group, and Common Item Mean and Variance Population Differences

	Mean Differences				Variance Differences			
	Observed	Forms	Group	Common	Observed	Forms	Group	Common
New Group 1								
E1	-0.6725	-0.6823	0.0097	0.0024	-2.7065	-2.5059	-0.2006	0.0048
M1	0.6898	0.6865	0.0033	0.0010	2.9341	2.7056	0.2284	0.0311
E2	-0.6557	-0.6771	0.0214	0.0051	-2.5214	-2.5517	0.0303	0.0015
M2	0.6766	0.6913	-0.0147	-0.0058	2.5631	2.5621	0.0010	-0.0184
New Group 2								
E1	0.0895	-0.6948	0.7844	0.1919	-1.7784	-1.9693	0.1909	-0.0519
M1	1.3600	0.7064	0.6536	0.2049	3.0911	2.3678	0.7233	0.0950
E2	1.2693	-0.7195	1.9888	0.4830	-1.6778	-1.1568	-0.5209	-0.2264
M2	2.3265	0.7315	1.5950	0.4979	2.7940	1.7794	1.0146	0.1543
New Group 3								
E1	-0.6216	-0.6553	0.0337	-0.0070	5.4715	-2.4170	7.8885	0.4658
M1	0.7178	0.6516	0.0661	0.0322	8.8479	2.7710	6.0769	0.5301
E2	-0.5324	-0.5979	0.0655	-0.0114	17.0678	-2.1175	19.1853	1.1331
M2	0.7561	0.6131	0.1430	0.0705	17.6444	2.9247	14.7197	1.2589
New Group 4								
E1	0.1573	-0.6545	0.8119	0.1872	6.0862	-1.8746	7.9608	0.4032
M1	1.3601	0.6760	0.6841	0.2241	8.9415	2.3636	6.5779	0.6023
E2	1.2506	-0.6069	1.8575	0.4253	18.1471	-0.9102	19.0572	0.9159
M2	2.2629	0.6431	1.6199	0.5274	17.6561	1.8828	15.7733	1.4050

Table A-9

Observed, Form, Group, and Common Item Mean and Variance Sample Differences

	Mean Differences				Variance Differences			
	Observed	Forms	Group	Common	Observed	Forms	Group	Common
New Group 1								
E1								
N=200	-0.6650	-0.6842	0.0192	0.0064	-2.5321	-2.5166	-0.0155	0.0141
N=400	-0.6701	-0.6810	0.0109	0.0008	-2.6591	-2.5637	-0.0954	0.0023
N=800	-0.6672	-0.6852	0.0180	0.0052	-2.8015	-2.4768	-0.3247	-0.0131
M1								
N=200	0.6975	0.6888	0.0087	0.0093	3.0009	2.7573	0.2436	0.0270
N=400	0.6873	0.6845	0.0028	0.0010	2.8556	2.7228	0.1328	0.0233
N=800	0.6893	0.6866	0.0027	0.0015	3.0667	2.7095	0.3572	0.0399
E2								
N=200	-0.6617	-0.6796	0.0179	0.0079	-2.5096	-2.4184	-0.0912	-0.0303
N=400	-0.6520	-0.6704	0.0184	0.0033	-2.4724	-2.6127	0.1403	0.0045
N=800	-0.6426	-0.6734	0.0308	0.0069	-2.5194	-2.5289	0.0095	0.0015
M2								
N=200	0.6616	0.7008	-0.0392	-0.0143	2.5908	2.5266	0.0642	-0.0246
N=400	0.6820	0.6923	-0.0103	-0.0061	2.3631	2.5255	-0.1624	-0.0372
N=800	0.6658	0.6909	-0.0251	-0.0078	2.5682	2.5935	-0.0253	-0.0274
New Group 2								
E1								
N=200	0.0834	-0.6951	0.7785	0.1913	-1.3998	-2.0607	0.6609	-0.0227
N=400	0.0832	-0.7004	0.7836	0.1885	-1.7717	-1.9367	0.1650	-0.0633
N=800	0.0838	-0.6951	0.7789	0.1923	-1.9291	-2.0624	0.1333	-0.0684
M1								
N=200	1.3648	0.7058	0.6590	0.2084	2.9813	2.3065	0.6748	0.0970
N=400	1.3719	0.7089	0.6630	0.2091	2.9998	2.3965	0.6033	0.0772
N=800	1.3512	0.7068	0.6444	0.2017	3.1620	2.3405	0.8215	0.1012
E2								
N=200	1.2621	-0.7165	1.9786	0.4847	-1.5666	-1.0474	-0.5192	-0.2403
N=400	1.2766	-0.7286	2.0052	0.4865	-1.7251	-1.1685	-0.5566	-0.2328
N=800	1.2706	-0.7196	1.9902	0.4833	-1.7860	-1.2071	-0.5789	-0.2283
M2								
N=200	2.3175	0.7293	1.5882	0.4959	2.7979	1.7564	1.0415	0.1463
N=400	2.3192	0.7342	1.5850	0.4945	2.6483	1.7435	0.9048	0.1461
N=800	2.3254	0.7324	1.5930	0.4987	2.7751	1.7248	1.0503	0.1544
New Group 3								
E1								
N=200	-0.6280	-0.6620	0.0340	-0.0103	5.9065	-2.3497	8.2562	0.5001

N=400	-0.6336	-0.6573	0.0237	-0.0103	5.5749	-2.4219	7.9968	0.4697
N=800	-0.6366	-0.6558	0.0192	-0.0085	5.4421	-2.3497	7.7918	0.4539
M1								
N=200	0.6956	0.6516	0.0440	0.0283	8.8783	2.6987	6.1796	0.5408
N=400	0.7340	0.6550	0.0790	0.0372	8.8299	2.7875	6.0424	0.5217
N=800	0.7237	0.6539	0.0698	0.0330	8.9370	2.7418	6.1952	0.5439
E2								
N=200	-0.5495	-0.6044	0.0549	-0.0155	16.9486	-2.1490	19.0976	1.1182
N=400	-0.5070	-0.5990	0.0920	-0.0060	17.0349	-2.1692	19.2041	1.1374
N=800	-0.5087	-0.5974	0.0887	-0.0068	17.0899	-2.0848	19.1747	1.1348
M2								
N=200	0.7519	0.6229	0.1290	0.0693	17.5485	2.7802	14.7683	1.2592
N=400	0.7616	0.6182	0.1434	0.0712	17.5114	2.8886	14.6228	1.2473
N=800	0.7587	0.6094	0.1493	0.0721	17.6346	2.9814	14.6532	1.2515
New Group 4								
E1								
N=200	0.1796	-0.6458	0.8254	0.1956	6.3001	-1.9247	8.2248	0.4301
N=400	0.1507	-0.6519	0.8026	0.1815	6.1078	-1.9785	8.0863	0.4153
N=800	0.1486	-0.6525	0.8011	0.1862	5.9862	-1.8168	7.8030	0.3814
M1								
N=200	1.3637	0.6717	0.6920	0.2306	8.8940	2.2899	6.6041	0.6132
N=400	1.3693	0.6795	0.6898	0.2270	8.8518	2.3579	6.4939	0.5901
N=800	1.3692	0.6746	0.6946	0.2255	9.0082	2.3320	6.6762	0.6129
E2								
N=200	1.2583	-0.6128	1.8711	0.4265	18.1145	-0.7706	18.8851	0.8973
N=400	1.2587	-0.6051	1.8638	0.4277	18.0566	-0.8608	18.9174	0.9067
N=800	1.2550	-0.6112	1.8662	0.4259	17.9946	-0.8985	18.8931	0.9104
M2								
N=200	2.2600	0.6409	1.6191	0.5295	17.6479	1.8182	15.8297	1.4053
N=400	2.2640	0.6415	1.6225	0.5295	17.6223	1.9223	15.7000	1.3997
N=800	2.2488	0.6423	1.6065	0.5244	17.6240	1.9037	15.7203	1.3915

Appendix B

Figures B-1 – B-7

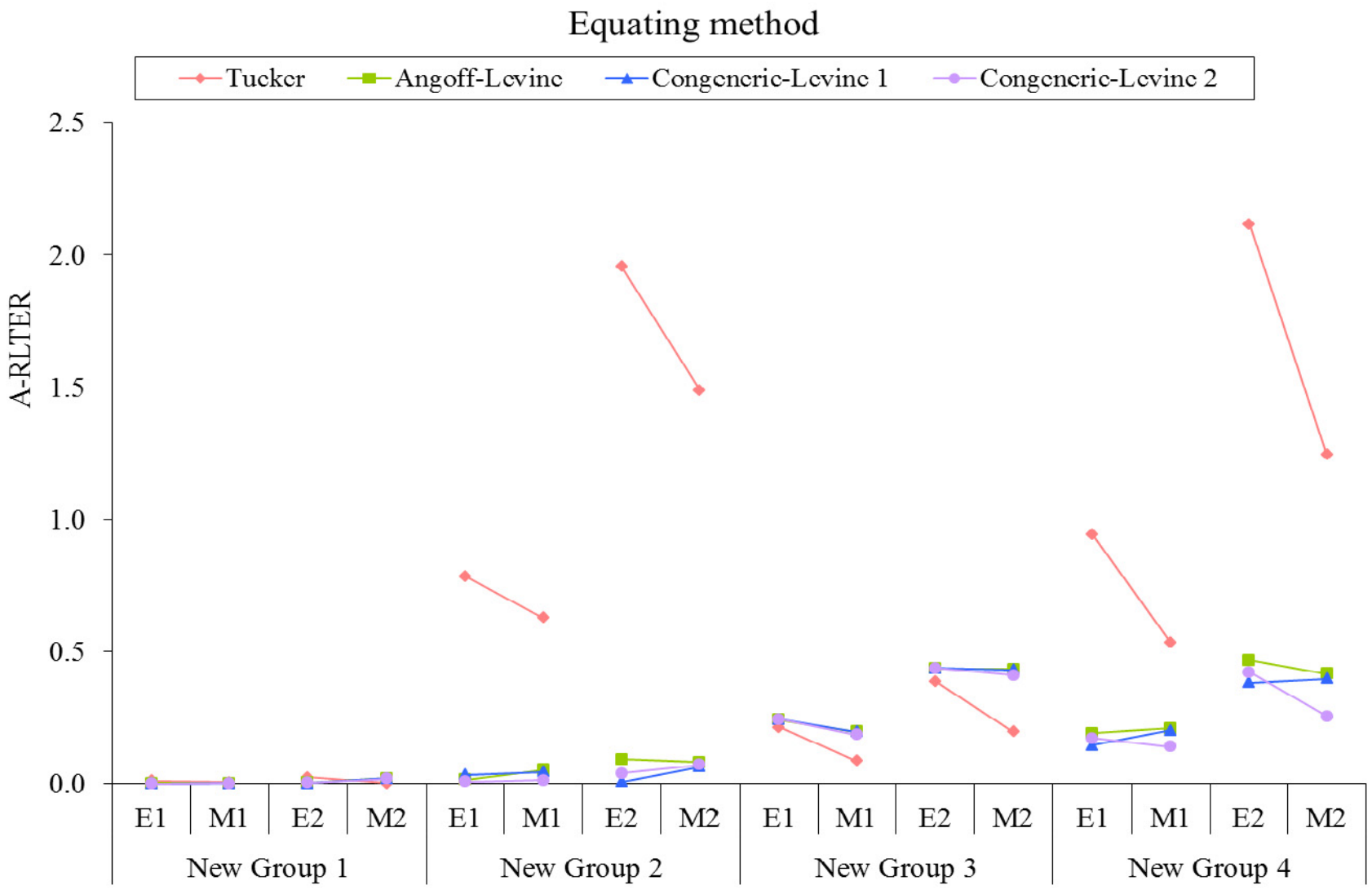


Figure B-1. Absolute relative error (A-RLTER) of the population mean indirect estimates, by equating method, simulation condition (E1, M1, E2, M2), and simulated population group (New Group 1 - New Group 4).

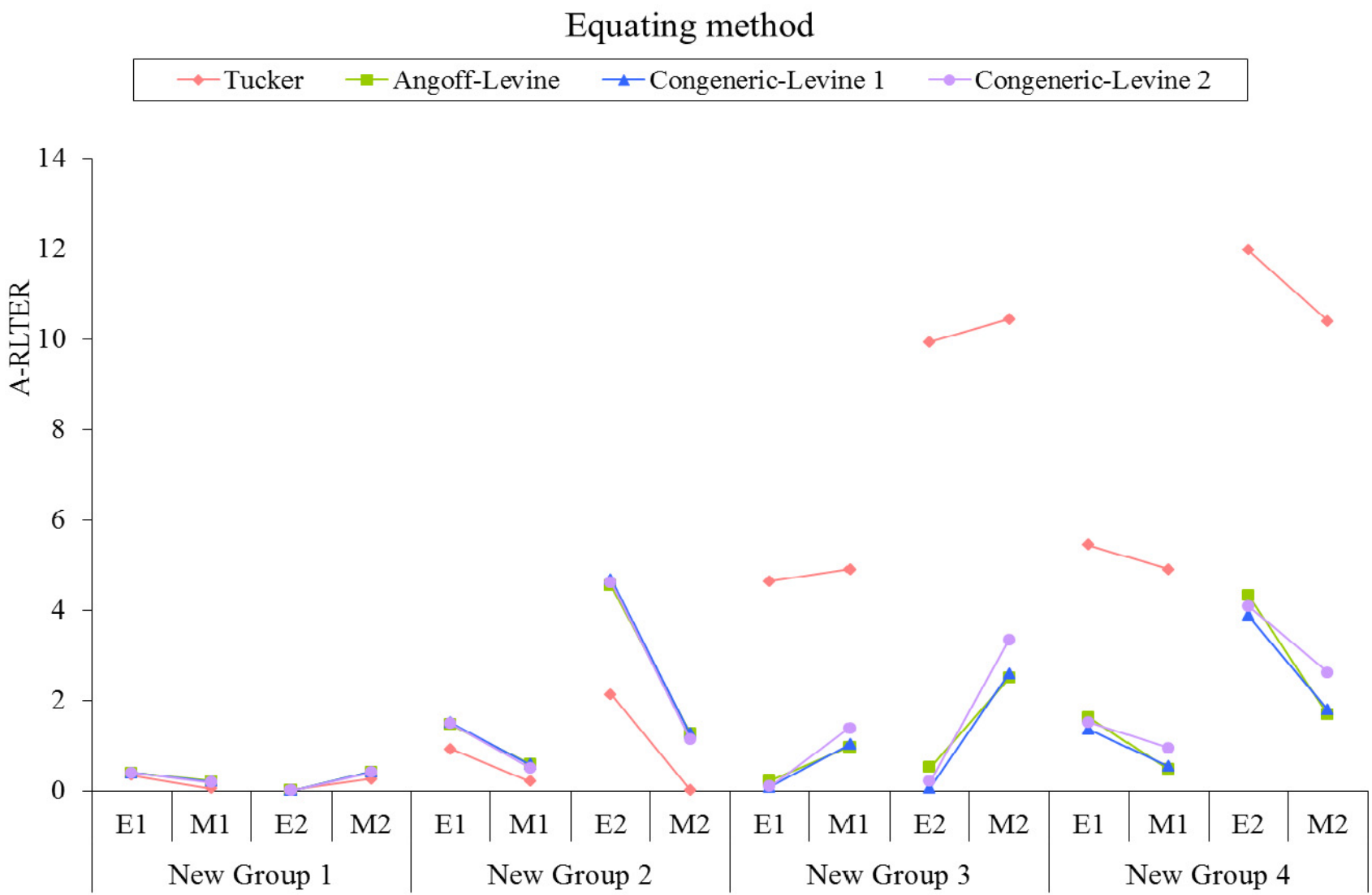


Figure B-2. Absolute relative error (A-RLTER) of the population variance indirect estimates, by equating method, simulation condition (E1, M1, E2, M2), and simulated population group (New Group 1 - New Group 4).

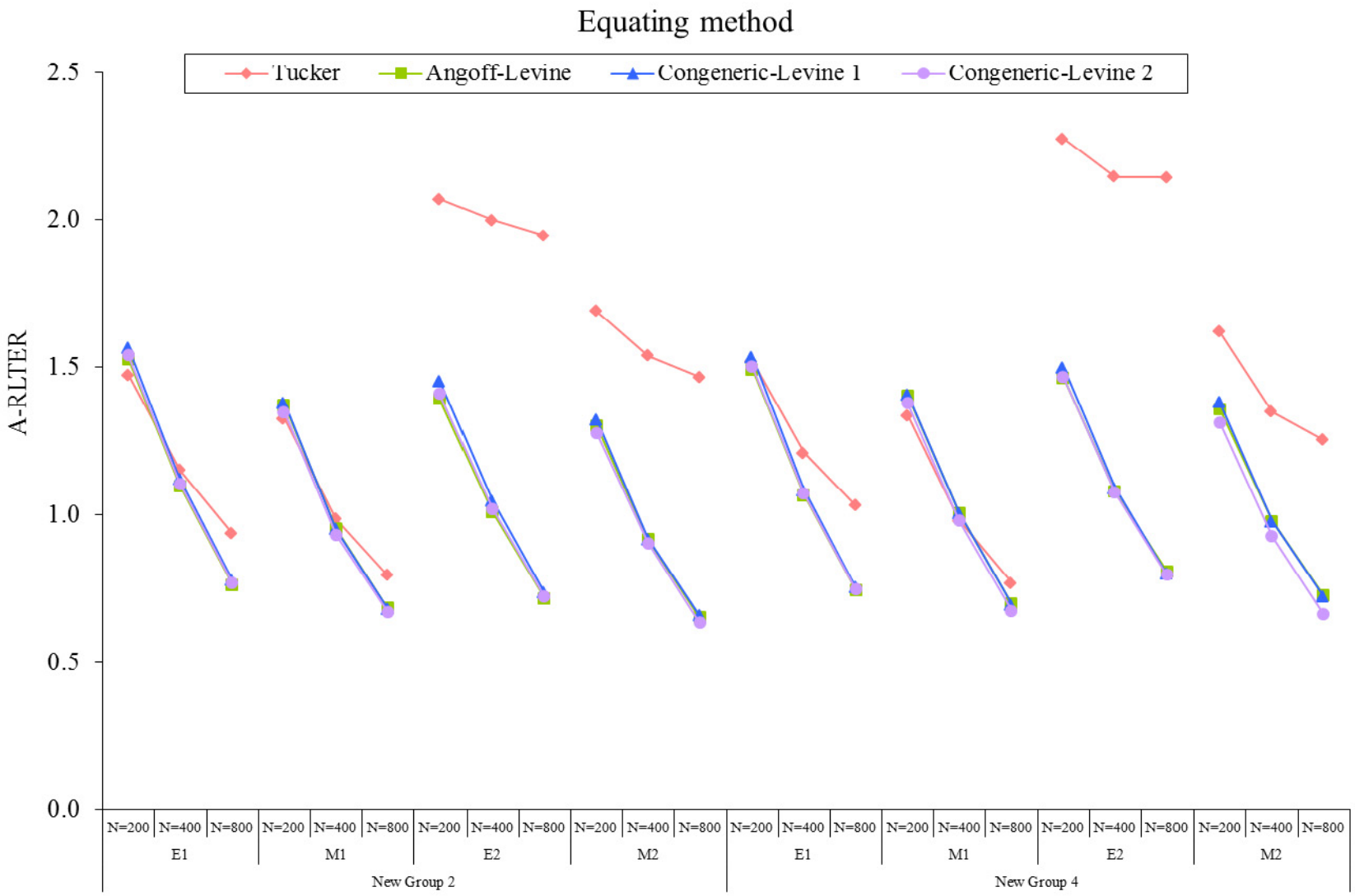


Figure B-3. Absolute relative error (A-RLTER) of the sample mean indirect estimates, by equating method, sample size (N), simulation condition (E1, M1, E2, M2), and simulated population group (New Group 2 and New Group 4).

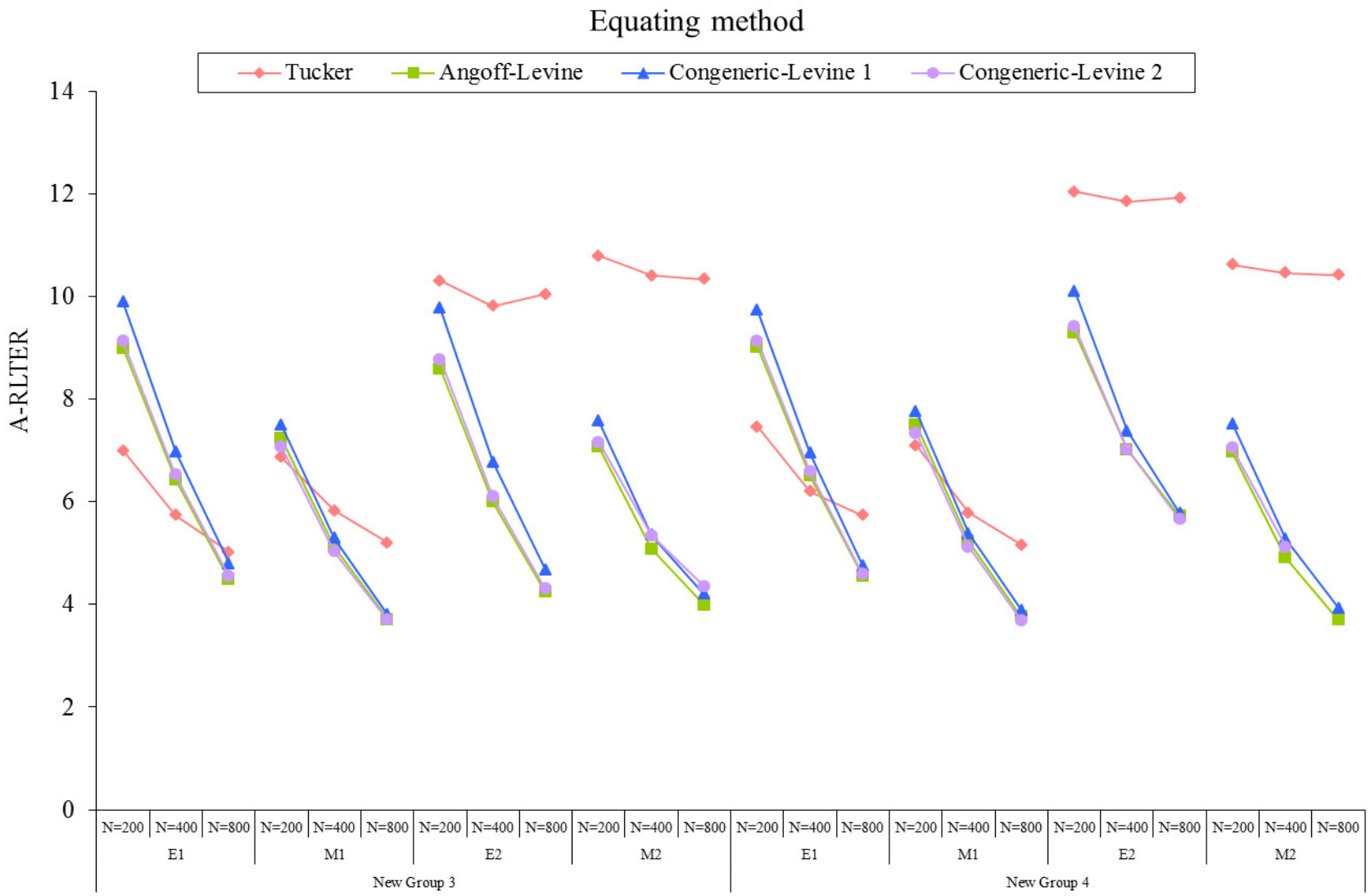


Figure B-4. Absolute relative error (A-RLTER) of the sample variance indirect estimates, by equating method, sample size (N), simulation condition (E1, M1, E2, M2), and simulated population group (New Group 3 and New Group 4).

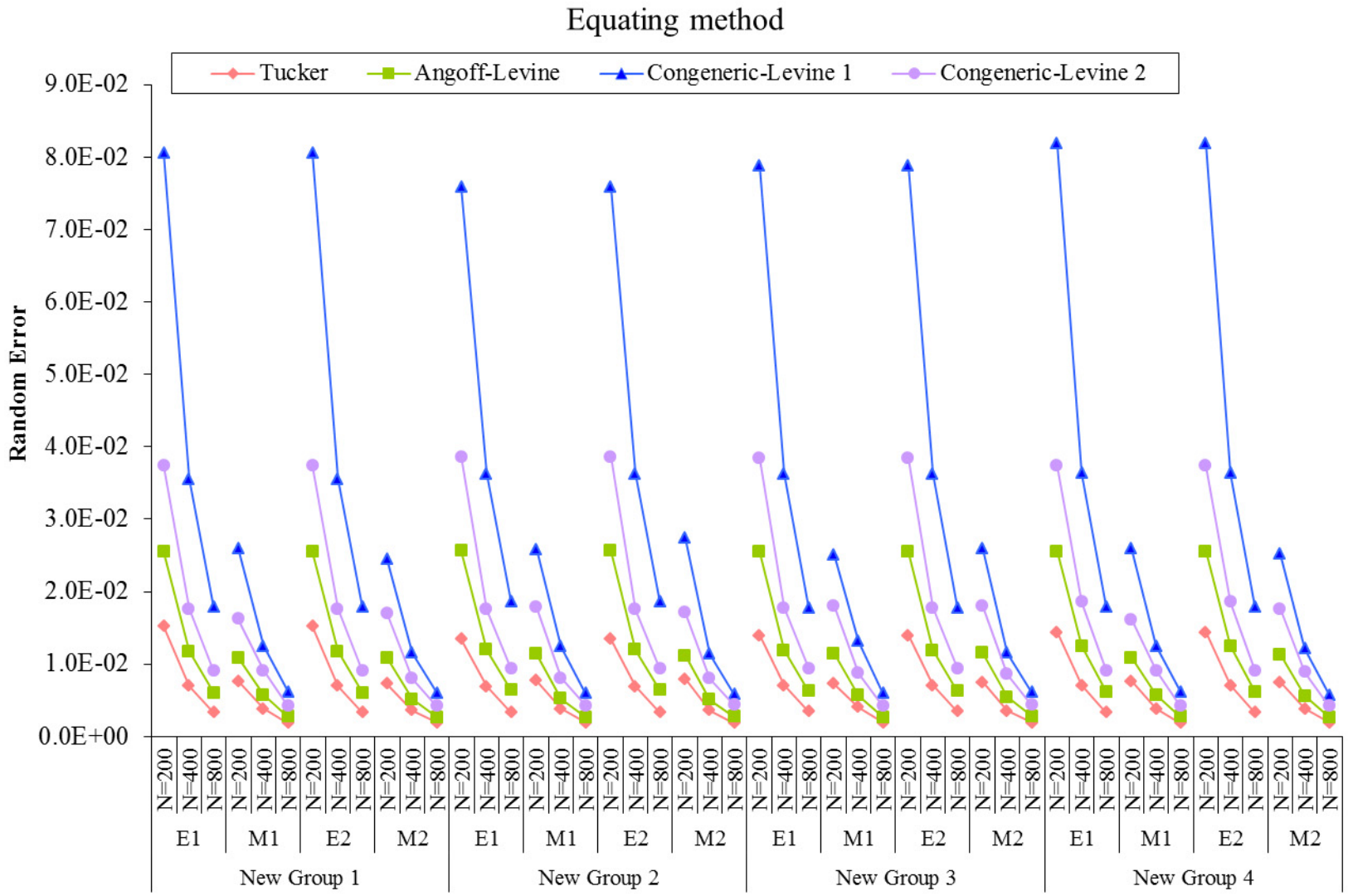


Figure B-5. Random error of the sample gamma coefficients by equating method, sample size (N), simulation condition (E1, M1, E2, M2), and simulated population group (New Group 1 - New Group 4).

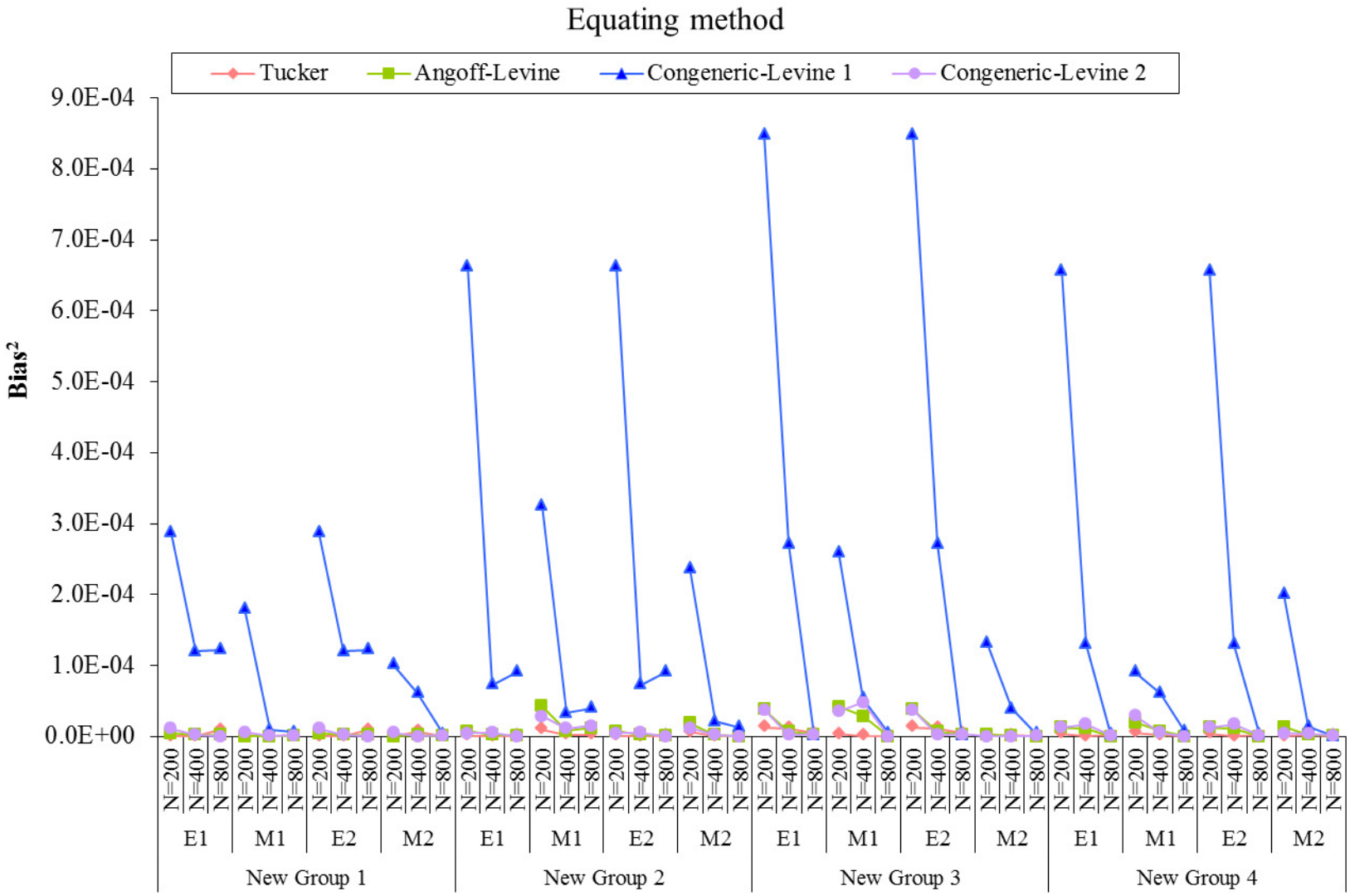


Figure B-6. Squared bias of the sample gamma coefficients by equating method, sample size (N), simulation condition (E1, M1, E2, M2), and simulated population group (New Group 1 - New Group 4).

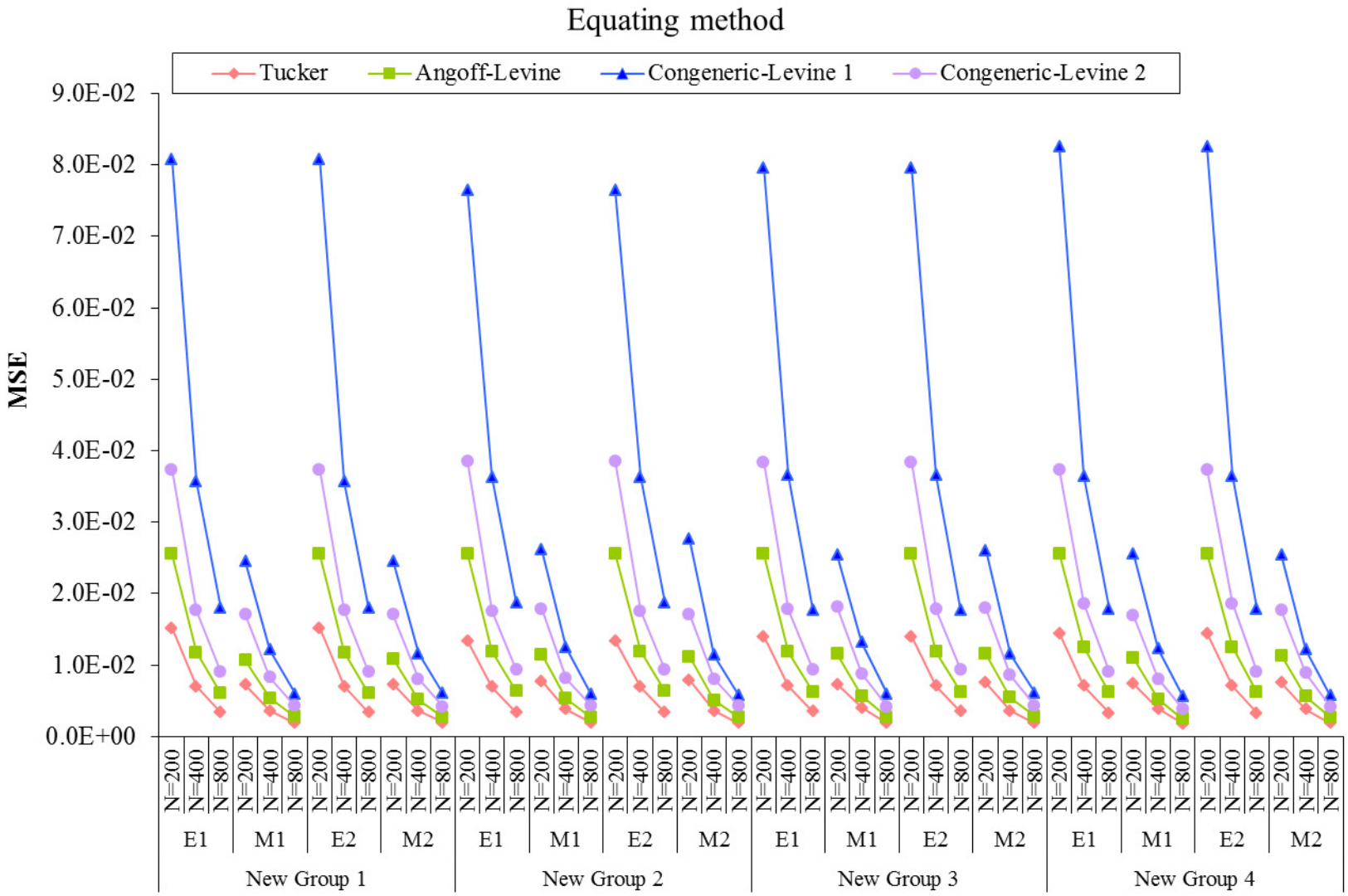


Figure B-7. Mean squared error (MSE) of the sample gamma coefficients by equating method, sample size (N), simulation condition (E1, M1, E2, M2), and simulated population group (New Group 1 - New Group 4).

05020213 0 Rev 1