

# Use of the estimated intraclass correlation for correcting differences in effect size by level

Soyeon Ahn · Nicolas D. Myers · Ying Jin

Published online: 30 September 2011  
© Psychonomic Society, Inc. 2011

**Abstract** In a meta-analysis of intervention or group comparison studies, researchers often encounter the circumstance in which the standardized mean differences ( $d$ -effect sizes) are computed at multiple levels (e.g., individual vs. cluster). Cluster-level  $d$ -effect sizes may be inflated and, thus, may need to be corrected using the intraclass correlation (ICC) before being combined with individual-level  $d$ -effect sizes. The ICC value, however, is seldom reported in primary studies and, thus, may need to be computed from other sources. This article proposes a method for estimating the ICC value from the reported standard deviations within a particular meta-analysis (i.e., estimated ICC) when an appropriate default ICC value (Hedges, 2009b) is unavailable. A series of simulations provided evidence that the proposed method yields an accurate and precise estimated ICC value, which can then be used for correct estimation of a  $d$ -effect size. The effects of other pertinent factors (e.g., number of studies) were also examined, followed by discussion of related limitations and future research in this area.

**Keywords** Intraclass correlation (ICC) · Meta-analysis · Standardized mean difference · Multilevel · Level of analysis

The standardized mean difference (also known as the  $d$ -effect size) has been widely used to evaluate the effects of interventions and treatments in the social sciences and education (Hedges, 2009b; Hedges & Hedberg, 2007). The  $d$ -effect size has been considered a scale-free index that quantifies the strength and direction of the intervention/

treatment effect or group mean difference (Cooper, Hedges, & Valentine, 2009). Thus, many researchers have combined  $d$ -effect sizes from multiple studies and have drawn a statistical inference about the overall intervention/treatment effect (e.g., Mol, Bus, & de Jong, 2009; Slavin, Lake, Chambers, Chueng, & Davis, 2009) or group mean difference (e.g., Swanson & Hsieh, 2009).

In a meta-analysis of studies in the social sciences and education, researchers often encounter the circumstance in which  $d$ -effect sizes are computed using summary statistics (i.e., means and standard deviations) originating from multiple levels (e.g., student vs. classroom or patient vs. clinic) across studies. For example, in a meta-analysis examining the effect of teachers' professional development programs on student mathematics achievement by Salinas (2010), two studies used aggregated data at the classroom level, while the rest of the studies were based on data at the student level.

Because individuals (e.g., students, clients) within the same cluster (e.g., classroom, counselors) are likely to be nonindependent, resulting in underestimated standard errors (Raudenbush & Bryk, 2002),  $d$ -effect sizes computed at the cluster level tend to be inflated, as compared with  $d$ -effect sizes computed at the individual level (Hedges, 2007, 2009a, b; What Works Clearinghouse [WWC], 2008). Thus,  $d$ -effect sizes from different levels should not be combined in a meta-analysis prior to the estimation of an overall effect size that accounts for the magnitude of nonindependence by level.

Such an issue often arises when studies included in a meta-analysis are based on *either* individual- or cluster-level data. Studies reporting findings from *both* cluster- and individual-level data could provide sufficient statistics to take account of dependency among samples within the same cluster when computing  $d$ -effect size. However, there are some cases in education and/or the social sciences

---

S. Ahn (✉) · N. D. Myers · Y. Jin  
University of Miami,  
Coral Gables, FL, USA  
e-mail: s.ahn@miami.edu

where providing findings from *both* cluster- and individual-level data is not feasible. For instance, a researcher using data gathered from a state may be restricted only to classroom data, rather than individual student data. In such a case, no information is provided at the individual level, which prevents the researcher from taking account of sample dependency when computing effect size. Clearly, in cases where only individual-level data are available, it would be incorrect for the researcher to rely on negatively biased standard errors that result from ignoring the nested nature of the data.

For handling such an issue, researchers have suggested correcting the  $d$ -effect size computed at the cluster-level ( $d_{clusters}$ ) and its associated variance ( $v_{(d_{clusters})}$ ) from the clusters using an intraclass correlation (ICC), which can ultimately be combined with the  $d$ -effect size computed at the individual level ( $d_{individuals}$ ). The formulas for correcting  $d_{clusters}$  that can be compatible with  $d_{individuals}$ , which is described in Hedges (2009b), are

$$d_{adjusted} = d_{clusters} * \sqrt{ICC}, \quad (1)$$

and

$$v_{(d_{adjusted})} = v_{(d_{clusters})} * ICC. \quad (2)$$

Consequently, the cluster-level  $d$ -effect size ( $d_{adjusted}$ ) and its variance ( $v_{(d_{adjusted})}$ ) adjusted by the ICC value are no longer biased and, thus, are compatible with the individual-level  $d$ -effect sizes.

The critical correcting factor in Eqs. 1 and 2 is the ICC value, which represents the degree to which observations in the same cluster are dependent due to shared variances (Hox, 2002; Kreft & de Leeuw, 1998). Simply, the ICC value ( $\rho$ ) is the proportion of cluster-level variance ( $\sigma_{clusters}^2$ ) to total variance ( $\sigma_{total}^2$ ) and is given by

$$\rho = \frac{\sigma_{clusters}^2}{\sigma_{total}^2} = \frac{\sigma_{clusters}^2}{\sigma_{clusters}^2 + \sigma_{individuals}^2}, \quad (3)$$

where  $\sigma_{individuals}^2$  is the individual-level variance.

However, the correcting factor given in Eqs. 1 and 2, the ICC value, is seldom reported in studies (Hedges, 2009b), which makes it difficult to correct for differences in the cluster-level  $d$ -effect sizes. As a resolution, some researchers have suggested imputing a plausible default ICC value from other sources. For instance, the WWC (2008) recommended using default ICC values of .20 for achievement outcomes and .10 for behavioral and attitudinal outcomes. These default values, which may not take ICC variation among different achievement variables or behavioral and attitudinal outcomes into account, were based on an analysis of the empirical literature in the field of education. Following the WWC's guideline, Salinas (2010) and Scher and O'Reilly (2009)

adjusted the cluster-level  $d$ -effect sizes using the default ICC value of .20 in their meta-analyses, both of which examined the effect of teachers' professional development programs on student achievement.

In addition, Hedges (2009b) proposed the computation of a *default ICC value* from a probability-based survey data set such as the National Assessment of Educational Progress (NAEP). For instance, Hedges and Hedberg (2007) have provided a summary of the ICC values for mathematics and reading computed from the national probability samples at different grade levels and in different regions of the United States (i.e., urban, suburban, rural). In their study, Hedges and Hedberg found that the average ICC across all grade levels was .22, which was computed using four existing national data sets (e.g., the Early Childhood Longitudinal Study and the Longitudinal Study of American Youth).

Although the method for extracting a default ICC value described above would be a reasonable option in certain circumstances, it is not practical for every context in which the clustering of observations occurs. For example, the suggested sources for the ICC value, such as the data set with the national probability samples or empirical studies reporting the ICC value, would not always be available. Moreover, even if they exist, these values may be limited for some specific subpopulations with unique characteristics. For instance, the default ICC value of .20 for student achievement suggested by WWC (2008) and Hedges (2009b) may not be appropriate for samples composed mostly of students from nonmajority ethnic backgrounds. For instance, Maerten-Rivera, Myers, Lee, and Penfield (2010) reported the ICC value of .13 for science achievements based on 23,854 fifth-grade students from 198 elementary schools in a large urban school district with a diverse student population. The ethnic background of the student population in their study consisted of 60% Hispanic, 28% African American, 11% White NonHispanic, and 1% Asian or Native American. Another instance in which a default ICC value of .20 may not be appropriate was provided by Myers, Feltz, Maier, Wolfe, and Reckase (2006), who provided evidence for an average ICC value of approximately .33 across physical education variables.

In resolving the practical limitations of the existing approach, the present study proposes using information from the included studies of any particular meta-analysis. Specifically, the ICC value, which is the proportion of between-cluster variance to total variance, can be estimated from variances/standard deviations and sample sizes that are typically reported in treatment/intervention or comparison studies. Thus, *the estimated ICC value* is not dependent on either the availability of external sources (e.g., a synthesis of probability-based survey national data sets) or the ICC value being provided in each study of a meta-analysis.

In the following sections, we first present how to estimate the ICC value from the reported standard deviations/variances

and sample sizes. Then, by using a Monte Carlo simulation, the performance of the ICC estimation was examined under different conditions, such as number of cluster ( $l$ ) and individual ( $m$ ) levels and the population ICC values ( $\rho$ ). Finally, the application of the ICC estimation in a series of hypothetical meta-analyses that varied by both study features [e.g., the population mean difference between control and treatment groups ( $\gamma_{trt} - \gamma_{ctr}$ ) and the number of studies included ( $l + m$ )] was investigated. Specifically, the overall effect-size estimator after correcting for differences in levels, using the estimated ICC value, was compared with one without the ICC correction and the other with a default ICC value (i.e.,  $\rho = .20$ ) correction in relation to other study features.

### Estimating the ICC value

Suppose that a hypothetical meta-analysis of  $l$  number of studies at the cluster level (e.g., classrooms) and  $m$  number of studies at the individual level (e.g., students) was undertaken. And from those studies, the variances/standard deviations and sample sizes for control and treatment groups were reported. From the  $l$  number of cluster- and the  $m$  number of individual-level variances/standard deviations for both control and treatment groups, let us define the  $l$ th cluster-level standard deviation (= square root of variance) and the  $m$ th individual-level standard deviation as  $SD_l$  ( $l = 1, 2, \dots, l - 1, l$ ) and  $SD_m$  ( $m = 1, 2, \dots, m - 1, m$ ), respectively. These individual-level and cluster-level standard deviations are closely related to the population ICC.

As is described in Raudenbush and Bryk (2002), the logarithmic transformed standard deviation is normally distributed with a sample mean of  $\log(SD) + [1/(2 * (n - 1))]$  and an approximate variance of  $1/(2 * (n - 1))$ . From the sampling distribution of the logarithmic transformed standard deviation, the weighted average cluster-level standard deviation can be estimated on the logarithmic scale by

$$\log(\hat{\sigma}_{cluster}) = \frac{\sum_{l=1}^l w_l s_l}{\sum_{l=1}^l w_l}, \tag{4}$$

where  $s_l$  is defined as  $\log(SD_l) + [1/(2 * (n_l - 1))]$  for the  $l$ th cluster-level study, where  $n_l$  is the number of clusters for  $l$ th cluster-level study;  $w_l$  is defined as  $1/v_l$ , where  $v_l$  is  $1/(2 * (n_l - 1))$ .

Because the estimate is on a logarithmic scale, it is transformed back to the original scale via

$$\hat{\sigma}_{cluster} = \exp(\log(\hat{\sigma}_{cluster})). \tag{5}$$

Following Eqs. 4 and 5, the cluster-level variances for both control ( $\hat{\sigma}_{cluster\_ctr}$ ) and treatment ( $\hat{\sigma}_{cluster\_trt}$ ) groups can be estimated.

Similarly, the weighted-average total standard deviation is estimated on the logarithmic scale:

$$\log(\hat{\sigma}_{total}) = \frac{\sum_{m=1}^m w_m s_m}{\sum_{m=1}^m w_m}, \tag{6}$$

where  $s_m$  is  $\log(SD_m) + [1/(2 * (n_m - 1))]$  for the  $m$ th individual-level study, where  $n_m$  is sample size for the  $m$ th individual-level study;  $w_m$  is defined as  $1/v_m$ , where  $v_m$  is  $1/(2 * (n_m - 1))$ . Then, the value is transformed back to the original scale, using

$$\hat{\sigma}_{total} = \exp(\log(\hat{\sigma}_{total})). \tag{7}$$

Again, using Eqs. 6 and 7, the total variances for control ( $\hat{\sigma}_{total\_ctr}$ ) and treatment ( $\hat{\sigma}_{total\_trt}$ ) groups would be estimated.

Finally, the ICC value can be estimated on the basis of the cluster-level standard deviation estimate and total standard deviation estimate,

$$\hat{\rho} = \frac{(\hat{\sigma}_{cluster\_ctr} + \hat{\sigma}_{cluster\_trt})^2}{(\hat{\sigma}_{total\_ctr} + \hat{\sigma}_{total\_trt})^2}, \tag{8}$$

where *ctr* and *trt* refer to control and treatment groups, respectively.

The ICC value estimated using Eq. 8, which is indeed a ratio of the estimated cluster-level variance to the estimated total variance, is independent of variations in the scale of the measures across studies.<sup>1</sup> Hence, the estimated scale-independent ICC value can be used to adjust differences on effect sizes by level, in cases in which the included studies in a meta-analysis employ widely varying scales of measures to represent the underlying construct of interest, which is often observed in practice.

### Simulation 1: ICC estimation

The first simulation examined the performance of the estimated ICC value on the basis of Eqs. 4–8 under

<sup>1</sup> In order to test whether or not the computed ICC value was sensitive to variation in the scale of measures, the authors of the present study conducted a small simulation study, in which the scale of measures was varied when computing the ICC value (Eq. 8). Mean bias and MSE values of the computed ICC value were .03 and < .00001, respectively, suggesting that the computed ICC was not affected by variation in the scale of measures. Such a result was found irrespective of any study features, including number of measures with different scales, population ICC value, sample size, and number of effect sizes. Full results of this simulation study are available by request to the first author.

different conditions. These conditions were determined by relevant factors (i.e.,  $l$ ,  $m$ ,  $n_c$ , and  $n_m$ ) that might affect the estimation of the ICC value. In this section, data generation and parameters used in the simulation were first discussed, followed by the presentation of simulation results.

#### Data generation

Using R (R Development Core Team, 2008) version 2.11.1, data for the population experimental and control groups, each having 30 individuals per cluster, were generated from the normal distributions with means (i.e.,  $\gamma_{ctr}$  and  $\gamma_{trt}$ ) and respective total variances (i.e.,  $\sigma_{ctr}^2$  and  $\sigma_{trt}^2$ ). Total variances were the sum of between-level variance (i.e.,  $\tau_{ctr}$  and  $\tau_{trt}$ ) and within-level variance (i.e.,  $\sigma_{within\_ctr}^2$  and  $\sigma_{within\_trt}^2$ ), which were determined by the population intraclass correlation (i.e.,  $\rho$ ). The between-level variances were generated from the normal distribution with a mean of 0 and respective between-level variances (i.e.,  $\tau_{ctr}$  and  $\tau_{trt}$ ). The within-level variances were generated from the normal distribution with a mean of 0 and respective within-level variances (i.e.,  $\sigma_{within\_ctr}^2$  and  $\sigma_{within\_trt}^2$ ).

From the generated population data for control and treatment groups, the  $l$  number of the  $v$  clusters and the  $m$  number of the  $w$  individuals were randomly sampled, and their standard deviations of the scores for the  $l$  number of clusters and the  $m$  number of individuals were computed for both treatment and control groups. From these observed standard deviations, the ICC value was estimated using Eqs. 4–8.

#### Choice of parameters

The relevant factors that would affect the estimation of the ICC value used in the simulation included the number of cluster levels ( $l$ ), cluster sample size ( $v$ ), the number of individual levels ( $m$ ), individual sample size ( $w$ ), the population mean difference between control and treatment groups ( $\gamma_{trt} - \gamma_{ctr}$ ), and the population ICC value ( $\rho$ ).

First, four population ICC values of .05, .15, .25, and .33 were used to represent the true population proportion of between-level variance to total variance. These were chosen to represent the small, small to medium, medium to large, and large between-level variances, respectively (Hox & Maas, 2001; Hox, Maas, & Brinkhuis, 2010). Second, three population mean differences between the control and treatment groups were set to 0.2, 0.5, and 0.7, indicating small, medium, and large group mean difference (Cohen, 1988). Last, the numbers of cluster-level and individual-level studies (i.e.,  $[l, m]$ ) were set to [5, 5], [10, 10], [20, 20], [30, 30], and [60, 60] with the fixed cluster size and

individual size of 30. In reality, the numbers of cluster and individual levels typically vary, and sample sizes often differ across studies as well. For simplicity, however, the fixed sample size of 30 equal for control and treatment groups were used to compute  $d$ -effect sizes for both cluster- and individual-level data.

In all, from the 12 populations (i.e.,  $[\gamma_{trt} - \gamma_{ctr}] = 0.2, 0.5, 0.7$  by  $\rho = .05, .15, .25, .33$ ), a total of 60 unique conditions (i.e.,  $[l, m] = [5, 5], [10, 10], [20, 20], [30, 30], [60, 60]$ ) were generated. These 60 different conditions were replicated 1,000 times, yielding 60,000 estimated ICC values.

#### Evaluation of estimator

Two criteria were used to evaluate the performance of the ICC estimation. One was the relative bias of the estimated ICC value (Hoogland & Boomsma, 1998), which is defined by

$$Bias(\hat{\rho}) = \frac{\hat{\rho} - \rho}{\rho}, \quad (9)$$

where  $\hat{\rho}$  is the mean of the ICC estimates across all replications for each condition, and  $\rho$  is the population ICC value. The absolute value of relative bias  $[Bias(\hat{\rho})]$  less than  $|-0.05|$  is considered to be an acceptable range of the relative bias value (Hoogland & Boomsma, 1998). And, the second index was the mean square error ( $MSE$ ) of the ICC estimates, which is computed as

$$MSE(\hat{\rho}) = [\hat{\rho} - \rho]^2 \times \text{var}(\hat{\rho}), \quad (10)$$

where  $E(\hat{\rho})$  is computed as the mean  $\hat{\rho}$  value and  $\text{var}(\hat{\rho})$  is the empirical variance of the  $\hat{\rho}$  values across the 1,000 replications for each combination.

#### Results

Table 1 displays the descriptive statistics of the relative bias and  $MSE$  values of the estimated ICC by all the conditions manipulated in the simulation. The average relative bias values of the estimated ICC ranged from 0.001 to 0.007, and the average  $MSE$  values were from 0.001 to 0.006, indicating that the ICC value was estimated accurately and precisely across three population mean differences and four population ICC values. In particular, the absolute value of the average relative bias values for all conditions was less than  $|-0.05|$ , indicating that both the relative bias values of the estimated ICC were in an acceptable range. The mean relative bias and  $MSE$  values of the estimated ICC were largest when the population ICC was set to .05 and .33, respectively. Because there were both negligible bias and

**Table 1** Relative bias and *MSE* value of the estimated ICC

Population Mean Difference	Population ICC Value	Relative Bias				<i>MSE</i>			
		<i>Min</i>	<i>Max</i>	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	<i>M</i>	<i>SD</i>
.2	.05	0.007	0.008	0.007	0.0005	0.001	0.003	0.002	0.001
	.15	0.002	0.003	0.002	0.0004	0.001	0.007	0.002	0.003
	.25	0.001	0.001	0.001	0.0002	0.0003	0.011	0.003	0.005
	.33	0.001	0.002	0.001	0.0004	0.001	0.023	0.006	0.010
.5	.05	0.006	0.007	0.006	0.0004	0.001	0.002	0.001	0.001
	.15	0.002	0.003	0.002	0.0003	0.001	0.006	0.002	0.002
	.25	0.001	0.002	0.002	0.0003	0.001	0.015	0.005	0.006
	.33	0.001	0.001	0.001	0.0004	0.000	0.020	0.005	0.008
.7	.05	0.006	0.007	0.007	0.0004	0.001	0.002	0.001	0.001
	.15	0.001	0.002	0.002	0.0003	0.001	0.006	0.002	0.002
	.25	0.001	0.002	0.001	0.0002	0.001	0.012	0.004	0.005
	.33	0.0005	0.001	0.001	0.0004	0.0003	0.021	0.005	0.009

similar precision across conditions (see Table 1), explaining the variance in these two outcomes was not undertaken. Simply, the estimated ICC was unbiased and relatively precise across conditions.

### Simulation 2: use of the estimated ICC in meta-analysis

The second simulation examined the application of the estimated ICC value to a set of hypothetical meta-analyses that varied by study features. Specifically, we used a simulation to examine the effect of incorporating the estimated ICC value from studies in meta-analyses of standardized mean differences (*ds*). We looked at the relative bias and *MSE* values of the estimators of mean effects, which varied by the ICC correction methods (i.e., no ICC correction, default ICC correction, and estimated ICC correction). We further compared the relative bias and *MSE* values of three overall mean effect-size estimators with different ICC correction methods in relation to the population mean difference, the number of studies ( $l + m$ ), and the ratio of the number of cluster-level *ds* ( $l$ ) to the number of individual-level *ds* ( $m$ ) [ $l : m$ ].

#### Data generation

From 12 population control and treatment groups that were generated in simulation 1 (i.e., 3 population mean differences of 0.2, 0.5, and  $0.7 \times 4$  population ICC values of .05, .15, .25, and .33), the set of hypothetical meta-analyses having  $l$  (i.e., number of studies using clusters) +  $m$  (i.e., number of studies using individuals) studies were created. From the included  $l + m$  studies, the cluster-level *d*-effect sizes were corrected using the estimated ICC value, and then the average variance-

weighted *d*-effect size<sup>2</sup> was computed and compared with the average weighted *d*-effect sizes without ICC correction and with a default ICC (i.e., .20) correction.

#### Choice of parameters

In addition to the two parameters used to generate the population data in simulation 1 (i.e., the population mean difference between control and treatment groups and the population ICC value), two other factors were added. One was the total number of studies included in the meta-analysis ( $l + m$ ), and the other was the proportion of the number of studies using clusters ( $l$ ) to the number of studies using individuals ( $m$ ) [ $l : m$ ].

*Number of studies included ( $l + m$ )* Ahn and Becker (2011) found that the number of studies included in meta-analysis ranged from 12 to 180, on the basis of a review of 71 meta-analyses published in the *Review of Educational Research* from 1990 to 2004 and the *Psychological Bulletin* from 1995 to 2004. Of 71 meta-analyses, Ahn and Becker showed that almost half of the studies included fewer than 50 studies in their studies. Therefore, two values of 12 (i.e., meta-analysis with the least numbers of studies) and 30 (i.e., meta-analysis with the average number of 12 and 50 studies) were set to total number of studies ( $l + m$ ).

*Ratio of  $l$  to the  $m$  ( $l : m$ )* Of the  $l + m$  number of the included studies, the following three sets of  $l : m$  [1:1 (i.e.,  $l = 6, m = 6$ ;  $l = 15, m = 15$ )], [2:1 (i.e.,  $l = 8, m = 4$ ;  $l = 20,$

<sup>2</sup> More details about the average variance-weighted effect size can be found in Cooper et al. (2009).

$m = 10$ ], and [1:2 (i.e.,  $l = 4, m = 8$ ;  $l = 10, m = 20$ )] were chosen as a ratio of cluster-level studies to individual-level studies. In reality, the ratio of cluster-level studies to individual-level studies would likely vary considerably; yet, for parsimony, these three patterns were chosen.

### Estimators

The overall variance-weighted mean effect-size ( $d_{est.}$ ) with the cluster-level  $d$ s adjusted by the estimated ICC value using Eqs. 4–8 were computed and compared with the overall variance-weighted mean effect size ( $d_{without}$ ) with no corrected cluster-level  $d$ s and the overall variance-weighted mean effect size with the cluster-level  $d$ s ( $d_{default}$ ) adjusted by the default ICC value of .20 (which was suggested by WWC, 2008, for educational outcomes). These three mean effect-size estimators were computed for each condition that varied by the population mean difference, the population ICC value, the number of studies included, and the ratio of cluster-level studies to individual-level studies.

To sum up, from 12 populations (i.e., 3 population mean differences of 0.2, 0.5, and  $0.7 \times 4$  population ICC values of .05, .15, .25, and .33), 72 unique conditions (i.e., 2 values of total number of studies (i.e., 12 and 30)  $\times$  3 values of a ratio of cluster-level studies to individual-level studies (i.e., [1:1], [1:2], and [2:1]) were generated. For each condition with 1,000 replications, three variance-weighted overall effect sizes (i.e.,  $d_{without}$ ,  $d_{default}$ , and  $d_{est.}$ ) were computed. Therefore, a total of 216,000 effect-size estimators were obtained from 72 different conditions, which were replicated 1,000 times.

### Evaluation of estimators

Two criteria—the relative bias and  $MSE$  values—were used to evaluate the overall effect-size estimators. Representing the overall  $d$ -effect size as  $\hat{\delta}$  and the population effect size as  $\delta$ , the relative bias and  $MSE$  values of the estimated  $d$ -effect size is computed by

$$Bias(\hat{\delta}) = \frac{\hat{\delta} - \delta}{\delta}, \quad (11)$$

and

$$MSE(\hat{\delta}) = [\hat{\delta} - \delta]^2 + \text{var}(\hat{\delta}), \quad (12)$$

where  $\hat{\delta}$  was computed as the mean  $\hat{\delta}$  value, and  $\text{var}(\hat{\delta})$  is the empirical variance of the  $\hat{\delta}$  values across the 1,000 replications for each combination. The absolute value of Bias ( $\hat{\delta}$ ) less than  $|-0.05|$  is considered to be an acceptable range of the relative bias value (Hoogland & Boomsma, 1998).

Descriptive statistics were first computed for the relative bias and  $MSE$  values of three estimators across the 72 conditions. Then two sets of ANOVAs were conducted to compare the relative bias and  $MSE$  of three variance-weighted overall effect sizes with different ICC correction methods (i.e.,  $d_{without}$ ,  $d_{default}$ , and  $d_{est.}$ ) in relation to the simulation parameters. The factors modeled in the ANOVAs included the population mean difference, the population ICC value, the total number of studies included, and the ratio of cluster-level studies to individual-level studies. The main effect of each parameter and the interaction effects with type of the ICC correction method were modeled in the ANOVAs as predictors of the relative bias and  $MSE$  values. In the ANOVAs, the significance level was set to .015, and the partial eta-squared value ( $\eta^2$ ), which is a relatively less sample-size-sensitive measure, was used to describe the impact of each predictor.

### Results

Table 2 displays the descriptive statistics of the relative bias and  $MSE$  values of three overall mean effect-size estimators ( $d_{est.}$ ,  $d_{without}$ ,  $d_{default}$ ) by the population mean difference and the population ICC value. As is shown in Table 2, the  $d_{est.}$  had average relative bias values ranging from 0.03 to 0.26 and average  $MSE$  values ranging from 0.002 to 0.003.  $d_{est.}$  had an average relative bias of less than  $|-0.05|$ , except in two conditions. The two exceptions showing a bias larger than  $|-0.05|$  were when the population value was negligible, in which the ICC correction by level was not necessary. Yet  $d_{est.}$  had approximately zero average  $MSE$  values across conditions. These results indicated that the estimated overall effect-size estimator with an estimated ICC correction was an accurate and precise one.

On the other hand, the estimated overall effect size without a correction ( $d_{without}$ ) had an average relative bias ranging from 0.27 to 1.71 and average  $MSE$  values ranging from 0.005 to 0.035. This result showed that the overall effect-size estimator without an ICC correction was biased and contained sizable errors. Similarly, the estimated overall effect size with a default ICC value of .20 correction ( $d_{default}$ ) had an average relative bias that was frequently bigger than an acceptable range of relative bias of  $|-0.05|$  (i.e.,  $\overline{Bias} = |-0.15|$  to  $|-0.75|$ ) and  $MSE$  values between 0.002 and 0.006, showing that  $d_{default}$  was also biased and inaccurate.

As is shown in the last three columns of Table 2, Cohen's  $d$ -effect size comparing mean differences between  $d_{without}$  and  $d_{est.}$  ranged from 3.43 to 4.62 in the absolute magnitude of bias values and from 1.26 to 2.43 in the  $MSE$  values, suggesting that the bias and  $MSE$  values of  $d_{without}$  are bigger than those of  $d_{est.}$  to a large degree. And Cohen's  $d$ -effect sizes for differences in the absolute magnitude of the bias values between  $d_{default}$  and  $d_{est.}$  ranged from 0.15 to

**Table 2** Relative bias and MSE value of the average *d*-effect size estimators

$\gamma_{trt}$	$\gamma_{ctrl}$	$\rho$	$d_{without}$				$d_{default}$				$d_{est}$				Cohen's <i>d</i> -Effect size						
			Min	Max	M	SD	Min	Max	M	SD	Min	Max	M	SD	Min	Max	M	SD	$d_{without}$ vs. $d_{default}$	$d_{without}$ vs. $d_{est}$	$d_{default}$ vs. $d_{est}$
Relative Bias																					
0.2	.05	.113	2.241	1.709	0.483	0.494	0.936	0.747	0.183	0.178	0.302	0.260	0.045	2.63	4.23	3.66					
			0.380	0.819	0.610	0.192	0.022	0.010	0.016	-0.037	-0.008	-0.019	0.010	4.40	4.35	-0.67					
			0.242	0.570	0.409	0.134	-0.045	-0.011	-0.024	0.012	-0.044	-0.009	-0.022	0.013	4.05	4.07	0.16				
0.5	.33	0.163	0.397	0.285	0.097	-0.175	-0.092	-0.141	0.032	-0.035	-0.008	-0.020	0.011	1.99	3.84	5.06					
			1.800	1.339	0.387	0.368	0.687	0.519	0.132	0.057	0.092	0.074	0.011	2.84	4.62	4.75					
			0.360	0.802	0.580	0.196	-0.017	0.023	0.003	0.017	-0.034	-0.011	-0.022	0.009	4.15	4.02	-1.40				
0.7	.33	0.163	0.392	0.278	0.098	-0.047	-0.037	-0.042	0.004	-0.034	-0.015	-0.024	0.007	3.58	3.78	3.16					
			1.448	1.063	0.336	0.226	0.483	0.356	0.111	-0.030	-0.021	-0.030	0.007	1.78	3.57	5.47					
			0.359	0.817	0.586	0.200	-0.006	0.040	0.018	0.018	-0.031	-0.011	-0.020	0.007	4.00	4.00	-0.15				
MSE	.25	0.221	0.519	0.367	0.132	-0.039	-0.028	-0.033	0.005	-0.030	-0.011	-0.021	0.008	3.58	3.70	1.80					
			0.160	0.390	0.274	0.102	-0.176	-0.112	-0.144	0.028	-0.035	-0.017	-0.026	0.007	1.74	3.43	5.78				
			0.2	.05	0.005	0.015	0.010	0.004	0.002	0.004	0.003	0.001	0.001	0.004	0.002	0.001	2.40	2.75	1.00		
0.003	0.009	0.006				0.002	0.001	0.004	0.002	0.001	0.001	0.004	0.002	0.001	2.53	2.53	0.00				
0.002	0.008	0.005				0.002	0.001	0.004	0.002	0.001	0.001	0.004	0.002	0.001	1.90	1.90	0.00				
0.5	.33	0.002	0.008	0.005	0.002	0.002	0.004	0.003	0.001	0.001	0.004	0.003	0.001	1.26	1.27	0.00					
			0.012	0.043	0.027	0.013	0.003	0.008	0.006	0.002	0.001	0.004	0.002	0.001	2.26	2.71	2.53				
			0.007	0.032	0.019	0.010	0.001	0.004	0.002	0.001	0.001	0.004	0.002	0.001	2.39	2.39	0.00				
0.7	.25	0.005	0.019	0.011	0.005	0.001	0.004	0.002	0.001	0.001	0.004	0.002	0.001	2.50	2.50	0.00					
			0.003	0.013	0.008	0.003	0.002	0.005	0.003	0.001	0.001	0.005	0.003	0.001	2.24	2.24	0.00				
			0.013	0.056	0.034	0.018	0.003	0.009	0.006	0.002	0.001	0.004	0.002	0.001	2.19	2.51	2.53				
MSE	.25	0.013	0.060	0.035	0.020	0.001	0.004	0.002	0.001	0.001	0.004	0.002	0.001	2.33	2.33	0.00					
			0.008	0.039	0.023	0.013	0.001	0.004	0.003	0.001	0.001	0.004	0.003	0.001	2.17	2.17	0.00				
			0.007	0.033	0.019	0.010	0.004	0.008	0.006	0.001	0.001	0.004	0.003	0.001	1.83	2.25	3.00				

Note. Overall effect-size estimator without ICC correction ( $d_{without}$ ); Overall effect-size estimator with a default ICC correction ( $d_{default}$ ); Overall effect-size estimator with an estimated ICC correction ( $d_{est}$ );  $[\gamma_{trt} - \gamma_{ctrl}] =$  the population mean difference treatment and control group;  $\rho =$  the population ICC value

5.78, indicating that the  $d_{est}$  was more accurate than  $d_{default}$ , yet the differences were small (with  $\rho$  of .25 for  $\gamma_{trt} - \gamma_{ctr}$  of .20,  $\rho$  of .15 for  $\gamma_{trt} - \gamma_{ctr}$  of .70) to large. For the  $MSE$  values, Cohen’s  $d$ -effect size between  $d_{default}$  and  $d_{est}$  ranged from 0 to 3.00, indicating that the  $d_{est}$  was more precise, except when  $\rho$  was either .15 or .25, whose  $MSE$  values were identical at 0.001. Overall, the  $d_{est}$  was the most accurate and precise, as compared with  $d_{without}$  and  $d_{default}$ , with exceptions on the identical  $MSE$  values of  $d_{est}$  and  $d_{default}$  when  $\rho$  was either .15 or .25

As is displayed in Table 3, the relative bias and  $MSE$  values of the estimators differed depending on the ICC correction method ( $F(2, 120) = 13,598.56, p < .01$ , for the relative bias value;  $F(2, 120) = 684.33, p < .01$ , for the  $MSE$  value). A post hoc comparison using Tukey’s adjustment showed that both relative bias and  $MSE$  values of  $d_{est}$  were lower than those of  $d_{without}$  ( $M_{diff} = -0.64, p < .01$ , for the relative bias value and  $M_{diff} = 0.015, p < .01$ , for the  $MSE$  value). And the relative bias and  $MSE$  values of  $d_{est}$  were lower than  $d_{default}$ , yet mean differences were not statistically significant ( $M_{diff} = -0.068, p = .13$ , for the relative bias value and  $M_{diff} = -0.001, p = .43$ , for the  $MSE$  value).

These three overall effect-size estimators (i.e.,  $d_{without}$ ,  $d_{default}$ , and  $d_{est}$ ) were next compared in relation to several study features manipulated in the simulation. These comparisons were based on two separate ANOVAs on the relative bias and  $MSE$  values of the overall mean effect-size estimators. Because the main purpose of each ANOVA was to compare three effect-size estimators in relation to other study features manipulated in the simulation, only the interaction effect related to type of ICC corrections (i.e., no

ICC correction, default ICC correction, and estimated ICC correction) with other study factors (i.e., the population mean difference, the population ICC value, the number of studies included, and ratio of number of cluster-levels to number of individual-levels) were modeled.

As is displayed in Table 3, results from ANOVAs indicated that there were significant three-way interactions of the ICC correction methods with the population mean difference, the population ICC value, and the ratio of numbers of clusters to individual levels on both relative bias and  $MSE$  values of the overall effect-size estimators. Because the higher-order interactions superseded main effects, only significant three-way interaction effects were investigated further.

Figure 1 displays the significant three-way interactions of the ICC correction methods (i.e., no ICC correction, default correction, and estimated ICC correction) with the relative bias values. The effect of the population mean difference on the relative bias of the overall effect-size estimators was relatively large, when the population ICC was set to .05 (see Fig. 1a). However, when the population ICC was set to either .15 or .25, mean differences in the relative biases of  $d_{default}$  and  $d_{est}$  were almost identical across all the population mean difference. However, mean differences in the relative bias values of  $d_{default}$  and  $d_{est}$  increased when the population ICC value was .33. In particular, when the population ICC value was equal to .33, the relative bias value of  $d_{default}$  was larger in a negative direction, as compared with that of  $d_{est}$ . As is shown in second column from the last of Table 2, Cohen’s  $d$ -effect size for differences in the absolute magnitude of the bias

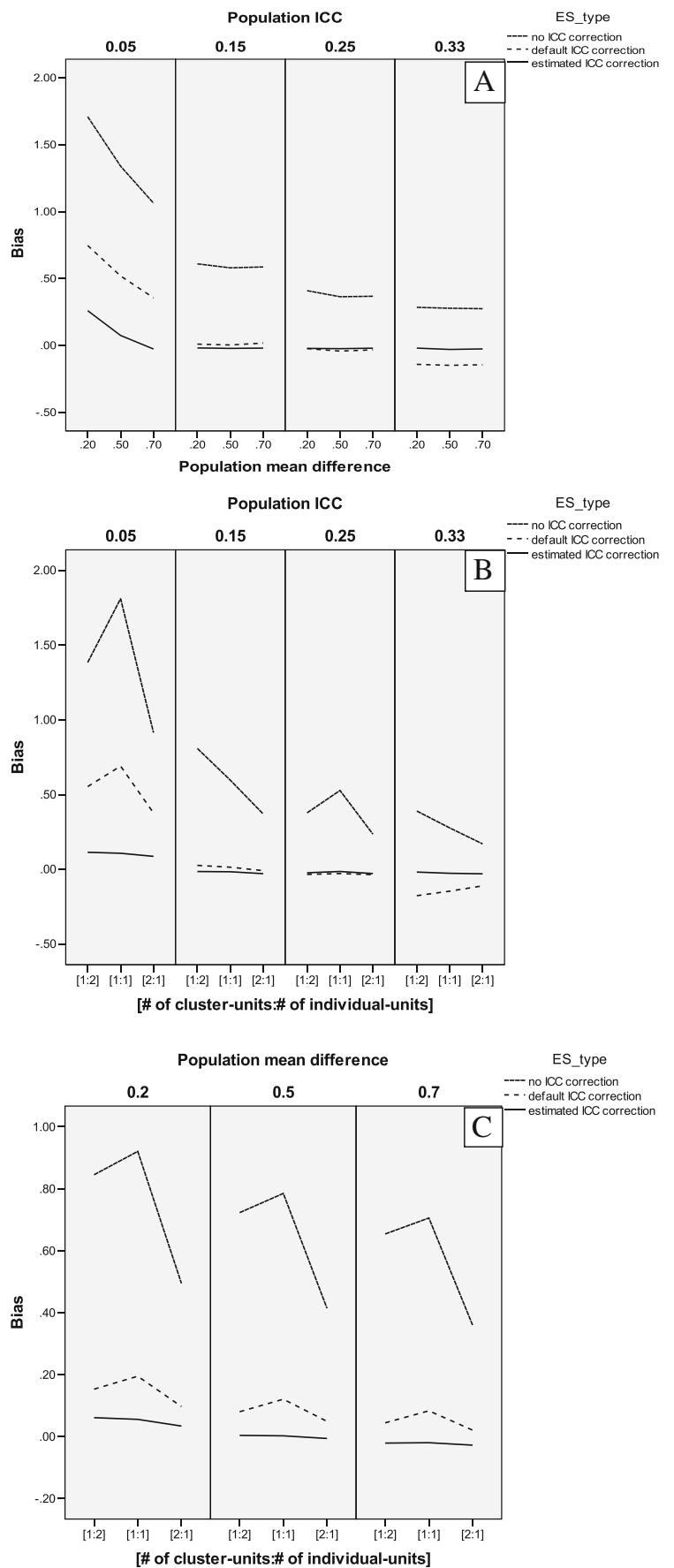
**Table 3** ANOVA results on the relative bias and MSE values of effect-size estimators

Source	Relative Bias					MSE				
	SS	df	MS	F	$\eta^2$	SS	df	MS	F	$\eta^2$
$[\gamma_{trt} - \gamma_{ctr}]$	0.51	2	0.26	389.80**	.87	0.002	2	0.001	145.34**	0.71
$\rho$	13.30	3	4.43	6773.50**	.99	0.0007	3	0.0002	33.94**	0.46
$[l + m]$	2.31E-05	1	2.31E-05	0.04	.00	0.0003	1	0.0003	43.15**	0.26
$[l : m]$	0.98	2	0.49	750.64**	.93	0.0008	2	0.0004	56.25**	0.48
$\rho_{correction}$	17.80	2	8.90	13598.56**	1.00	0.0095	2	0.005	684.33**	0.92
$[\gamma_{trt} - \gamma_{ctr}] * \rho * \rho_{correction}$	1.40	18	0.08	118.98**	.95	0.0006	18	3.39E-05	4.90**	0.42
$[l : m] * \rho * \rho_{correction}$	1.64	18	0.09	139.25**	.95	0.0018	18	0.0001	14.75**	0.69
$[l + m] * \rho * \rho_{correction}$	0.001	9	0.0002	0.24	.02	1.97E-06	9	2.19E-07	.03	0.002
$[\gamma_{trt} - \gamma_{ctr}] * [l : m] * \rho_{correction}$	0.02	12	0.0016	2.40**	.19	0.001	12	9.29E-05	13.40**	0.57
$[\gamma_{trt} - \gamma_{ctr}] * [l + m] * \rho_{correction}$	0.001	6	0.0002	0.28	.01	3.95E-07	6	6.58E-08	.01	0.0005
$[l + m] * [l : m] * \rho_{correction}$	0.0003	6	4.23E-05	0.06	.003	3.04E-06	6	5.06E-07	.07	0.004
Error	0.08	120	0.0007			0.0008	120	6.93E-06		
Total	42.03	215				0.02	215			

Note. \*\*  $p < .01$ ;  $[\gamma_{trt} - \gamma_{ctr}]$  = the population mean difference between treatment and control groups;  $\rho$  = the population ICC value;  $[l + m]$  = the number of studies;  $[l : m]$  = [# of cluster-levels : # of individual-levels];  $\rho_{correction}$  = type of the ICC correction method



**Fig. 1** Relative bias of the overall effect size by study features



values between  $d_{default}$  and  $d_{est}$  were all over 5, indicating that  $d_{est}$  has lower bias values, as compared with  $d_{default}$ .

The relative bias value of  $d_{est}$  was consistent across the four population ICC values regardless of the ratio of the number of cluster-level studies to the number of individual-level studies. And across all population ICC values, the differences between  $d_{est}$  and  $d_{default}$  in the relative bias were negligible. However, the effects of different ratios of cluster and individual levels on the relative bias value of  $d_{without}$  and  $d_{default}$  were different across four ICC values. In particular, the overall mean-effect estimators (i.e.,  $d_{without}$  and  $d_{default}$ ) with more individual-level studies were more accurate (see Fig. 1b).

In addition, the effects of different proportions of cluster levels and individual levels on the relative bias of  $d_{est}$  were fairly consistent for all population mean difference (see Fig. 1c). However, as is shown in Fig. 1c, both  $d_{default}$  and  $d_{without}$  had smaller relative bias values when the individual levels were twice the cluster-levels, while these had larger relative bias values with the same numbers of individual and cluster levels included.

Figure 2 presents the significant three-way interactions of the ICC correction methods on the  $MSE$  value of the overall effect-size estimators. As is shown in Fig. 2a, the  $MSE$  value of  $d_{est}$  was consistently low (almost zero) regardless of the population mean difference and the population ICC value. However, the  $MSE$  values of  $d_{default}$  and  $d_{without}$  differed considerably by both the population ICC value and the population mean difference. Specifically,  $d_{default}$  had almost zero  $MSE$  values when the population ICC was either .15 or .25, but its  $MSE$  values increased as the population mean difference got bigger only with the population ICC value of .05 and .33. For the  $MSE$  values, Cohen's  $d$ -effect size between  $d_{default}$  and  $d_{est}$  was over 1 when the population ICC values was either .05 or .33, indicating that the  $d_{est}$  was more precise, except when  $\rho$  was either .15 or .25. And the differences in the  $MSE$  value of  $d_{without}$  by the population mean difference became smaller when the population ICC became larger (see Fig. 2a).

As is displayed in Fig. 2b, differences in  $MSE$  values of  $d_{default}$  and  $d_{without}$  by different ratios of cluster and individual levels were bigger as the population mean difference became larger. However, the  $MSE$  values of  $d_{est}$  were consistently low (almost zero) regardless of the population mean difference or the population ICC. And having more cluster levels yielded slightly bigger  $MSE$  values of both  $d_{default}$  and  $d_{est}$ , yet produced smaller  $MSE$  values of  $d_{without}$ .

Lastly, having a different ratio of the number of cluster-level studies to the number of individual-level studies on the  $MSE$  values of  $d_{est}$  was consistent across four population ICC values. However, the effect of different ratios of cluster and individual levels on the  $MSE$  values of

$d_{without}$  were not consistent across four ICC values, showing that having more individual-level studies made the  $MSE$  values of  $d_{without}$  smaller (see Fig. 2c).

## General discussion

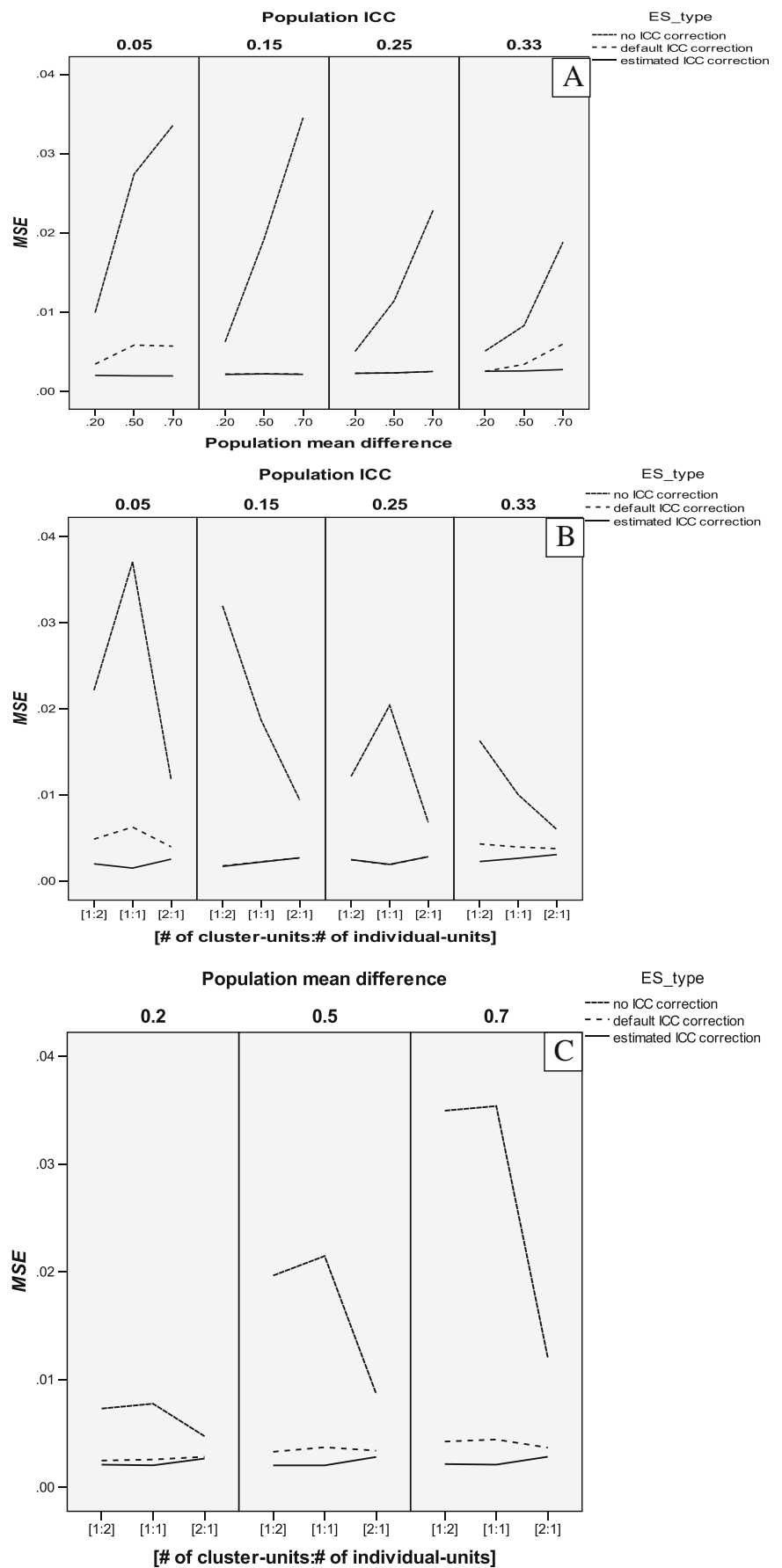
The present study focused on the central issue of synthesizing  $d$ -effect sizes originating from different levels (e.g., students and classrooms). This issue often arises when the included studies examine the effect of an intervention or treatment that can be implemented at different levels. Also, a mixture of studies providing  $d$ -effect sizes from both cluster-level and individual-level data can be found when comparing naturally occurring groups. For instance, the effect of teachers' having a bachelor degree (BA) in mathematics on student math achievement can be studied by either comparing mean scores from classrooms of teachers with BAs with those of teachers without BAs or comparing students mean scores of two teacher groups.

Although the danger of drawing inferences about individual behavior on the basis of cluster data (WWC, 2008), which is referred to as the *fallacy of ecological inference* (Robinson, 1950), has long been discussed, many researchers still utilize aggregated data, and many meta-analyst may include these studies in their syntheses. The use of aggregated data could be done for the convenience of the researchers and/or because the meaning of the effect is similar across levels. There are also some situations in which the individual-level group assignments are not feasible for political or practical reasons (Hedges & Hedberg, 2007) or the individual-level data are not accessible (e.g., the restricted use of individual-level data). In all of these cases, the proposed method may assist in combining studies originating at different levels.

The correction of the cluster-level  $d$ -effect sizes using the ICC value is necessary because the cluster-level  $d$ -effect sizes are upwardly biased. However, it can be challenging to obtain the ICC value because the studies using clustered data do not often report it (Hedges, 2009b). Although the existing recommendations for extracting a default ICC value, particularly on the basis of empirical searches of previous studies or data sets using the national-level probability samples, might be reasonable in some cases, practical limitations remain. In resolving such limitations, the present study proposes incorporating the estimated ICC from standard deviations/variances that are often reported (unlike the ICC value) in the included studies for any particular meta-analysis.

As was shown in the first simulation, the ICC value was accurately and precisely estimated from the standard deviations, with the average relative bias less than an acceptable range of  $[-.05]$  and  $MSE$  values approximately

**Fig. 2** MSE values of the overall effect size by study features



close to zero. Also, the ICC value was not sensitive to variation in the scale of the measures that represented the same underlying construct of interest. The accuracy and precision of the ICC estimation were consistent irrespective of any study features. In other words, the proposed model produced the accurate and precise estimation of the ICC value, which eventually affects the correct estimation of the overall effect size in a meta-analysis.

In general, the estimated overall effect size after correcting the cluster-level  $d$ -effect size using the estimated ICC value was unbiased and accurate, having average relative bias less than  $|-0.05|$  and  $MSE$  values close to zero. Specifically, the overall effect-size estimator with the estimated ICC correction had significantly lower mean relative bias and  $MSE$  values, as compared with no ICC correction. Even though the mean difference in the relative bias and  $MSE$  values of the effect-size estimators was not statistically significant, the overall effect size with the estimated ICC correction was lower than that with a default ICC correction ( $M_{diff} = -0.068$  for the relative bias value and  $M_{diff} = -0.001$  for the  $MSE$  value).

The relative bias and  $MSE$  values of the overall effect-size estimators with different ICC corrections varied depending on the population mean difference, the population ICC value, and the ratio of cluster-level studies to individual-level studies. Specifically, the relative bias and  $MSE$  values of the overall effect size with the estimated ICC correction were similar to that of the overall effect size with a default ICC correction, when the population ICC value was set to either .15 or .25. This makes sense because the computed overall effect size with a default ICC correction was based on the cluster-level  $d$ -effect sizes corrected by a default ICC value of .20. However, relative bias and  $MSE$  values of the overall effect size with a default ICC correction were bigger than those with the estimated ICC correction when the population ICC value was set to either .05 or .33.

The advantage of the proposed method largely comes from the accuracy and precision of the ICC estimation, which does not appear to be sensitive to variation in the scale of measures, and leads to the correct estimation of the overall effect size in a meta-analysis. In addition, the ease of utilizing the proposed method for any contexts of interest is quite appealing. Specifically, the estimation of both between and total variances for the ICC computation is a simple application of the regular meta-analytic procedure. Therefore, any meta-analysts can easily extract the ICC value from the reported standard deviations and sample sizes. Moreover, the estimated ICC value from studies would be better to represent the unique characteristics of the population whose data are nested in nature.

In spite of these practical advantages of the proposed method, there are a few methodological concerns. First, the present study assumes that standard deviations/variances are

from the same population, and thus the fixed-effect model is used to estimate the between and within variances. In cases where the fixed effect of standard deviation is suspicious, the random-effects model should be used to estimate the average variances (Raudenbush, 1994). Second, the proposed method assumes that standard deviations and sample sizes are available from the included studies. Such concern would be relatively trivial, since most intervention and comparison studies report summary statistics. Lastly, the parameters used in simulations 1 and 2 might be too hypothetical to represent the reality. For instance, in reality, the numbers of cluster and individual levels vary considerably, and sample sizes often differ across studies as well.

In spite of these methodological concerns, the proposed method appears to be both practical and applicable to many contexts of research synthesis. Also, the estimated overall effect size after correcting cluster-level effect sizes is unbiased and accurate, which is due mainly to the correctly estimated ICC value incorporated into the overall effect-size estimation. However, it should be reemphasized that the practicality of the proposed method is dependent on the primary studies providing the necessary information for estimating the ICC value, particularly standard deviations and sample sizes.

## References

- Ahn, S., & Becker, B. J. (2011). *Incorporating quality scores in meta-analyses*. Manuscript submitted for publication.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Cooper, H., Hedges, L., & Valentine, J. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.
- Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, 32, 151–179.
- Hedges, L. V. (2009a). Adjusting a significance test for clustering in designs with two levels of nesting. *Journal of Educational and Behavioral Statistics*, 34, 464–490.
- Hedges, L. V. (2009b). Effect sizes in nested designs. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 337–355). New York: Russell Sage Foundation.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation for planning group randomized experiments in education. *Educational Evaluation and Policy analysis*, 29, 60–87.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, 26, 329–367.
- Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157–174.
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64, 157–170.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

- Maerten-Rivera, J., Myers, N. D., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education, 94*, 937–962.
- Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research, 79*, 979–1007.
- Myers, N. D., Feltz, D. L., Maier, K. S., Wolfe, E. W., & Reckase, M. D. (2006). Athletes' evaluations of their head coach's coaching competency. *Research Quarterly for Exercise and Sport, 77*, 111–121.
- R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Raudenbush, S. W. (1994). Random-effects model. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–323). New York: Russell Sage Foundation.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Robinson, W. S. (1950). Ecological correlation and the behavior of individuals. *American Sociological Review, 15*, 351–357.
- Salinas, A. (2010). *Investing in teachers: What focus of professional development lead to the highest student gains in mathematics achievement?*. Unpublished doctoral dissertation, University of Miami.
- Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teacher: What do we really know? *Journal of Research on Educational Effectiveness, 2*, 209–249.
- Slavin, R. E., Lake, C., Chambers, B., Chueng, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*, 1391–1466.
- Swanson, H. L., & Hsieh, C.-J. (2009). Reading disabilities in adults: A selective meta-analysis of literature. *Review of Educational Research, 79*, 1362–1390.
- What Works Clearinghouse. (2008). *Procedures and standards handbook (version 2.0)*. Retrieved June 4, 2010, from [http://ies.ed.gov/ncee/wwc/pdf/wwc\\_procedures\\_v2\\_standards\\_handbook.pdf](http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf)