Journal of Modern Applied Statistical Methods
https://jmasm.com

Article

Comparing Three Approaches to Handling Dependency: A Case Study of a Meta-Analysis of Writing Self-Efficacy and Writing Proficiency

Ting Sun [1,*], Chuang Wang [2] and Richard G. Lambert [2]

[1] Department of Surgery, University of Utah Spencer Fox Eccles School of Medicine, 30 N Mario Cappechi Drive, Salt Lake City, UT 84132, USA
[2] Department of Educational Leadership, University of North Carolina at Charlotte Cato College of Education, Charlotte, NC 28223, USA
* Correspondence: ting.sun@hsc.utah.edu; Tel.: +(704)-858-4943

How To Cite: Sun, T.; Wang, C.; Lambert, R.G. Comparing Three Approaches to Handling Dependency: A Case Study of a Meta-Analysis of Writing Self-Efficacy and Writing Proficiency. *Journal of Modern Applied Statistical Methods* **2025**, *24*(2), 5. https://doi.org/10.53941/jmasm.2025.100005

Abstract: This study compared three approaches (i.e., averaging within-study effect sizes, three-level meta-analysis, and robust variance estimation) to handle dependent correlational effect sizes in conducting a meta-analysis. Data were from a meta-analytic study examining the relationship between writing self-efficacy and writing proficiency. To examine the differences in the performance of the three approaches, seven conditions were created by the number of studies and the number of effect sizes per study. While all three approaches produced similar results in the average effect size and standard error, the averaging approach had much smaller variance estimates. The patterns were basically consistent across different conditions. This study informs meta-analysts of appropriate procedures in handling the dependent effect sizes.

Keywords: dependency; effect size; averaging; three-level meta-analysis; robust variance estimation

1. Introduction

In social and behavioral sciences, it is common to see inconsistencies in statistical results yielded by similar studies. These discrepancies may come from either random sampling errors or heterogeneity of research designs across studies [1]. A meta-analysis provides an opportunity to make a scientific and systematic synthesis of empirical studies, which would help in the estimation of overall average effect sizes in one domain [2]. According to Borenstein et al. [1], effect sizes are indices measuring the effectiveness of a treatment (e.g., odds ratio), the effectiveness of an intervention (e.g., standardized mean difference), or the relationship between two variables (e.g., correlation coefficient). A meta-analysis can not only estimate overall effect sizes but also make projections about the variation of these effect sizes across studies. This would give us information and implications about what may capture the true variation among the effect sizes.

Given the appealing features of meta-analyses, growing attention has been directed to the validity of meta-analysis results [3]. One of the challenges to validity arises when synthesizing primary studies that report multiple dependent effect sizes. This is a common issue in conducting meta-analytic studies in social sciences [4]. Cheung and Chan [5] found that 31 out of 49 meta-analyses published in the Journal of Applied Psychology from 1991 to 2001 encountered dependent within-study effect sizes. Ahn et al. [3] reviewed meta-analyses in education published from 2000 to 2010 and found approximately two-thirds of them were confronted with the dependency issue.

Tanner-Smith and Tipton [6] classified dependency in meta-analyses into two types: hierarchical dependency and correlated dependency. Hierarchical dependency occurs when one study reports effect sizes from multiple

samples or the same research teams report effect sizes from multiple studies. Correlated dependency in a meta-analysis may result from multiple time points, multiple measures, or multiple comparisons within a study [1,7]. For example, Hashemnejad et al. [8] reported effect sizes on the relationship between self-efficacy and writing outcomes at three different points in time at a one-week interval. Perin et al. [9] reported eight effect sizes calculated from multiple writing outcome measures (e.g., analytic quality, holistic quality). Multiple comparisons within a study can be exemplified by Shintani and Aubrey's study [10], which reported effect sizes calculated from two treatment groups (i.e., synchronous corrective feedback and asynchronous corrective feedback) and one control group (i.e., no corrective feedback). These effect sizes are correlated since they are yielded with the same designs, conducted by the same researchers, measured by the same instruments, or calculated with the same samples [2].

Traditional meta-analysis procedure (i.e., univariate meta-analysis) is not appropriate in synthesizing studies reporting dependent effect sizes because the univariate meta-analysis assumes the independence of effect sizes [1]. Since effect sizes within studies are correlated with each other, ignoring their dependency would have the risk of underestimating the standard error of the overall effect size estimates and committing a Type I error [1,4]. In addition, the estimates of the overall average effect size would be biased towards the studies having more effects because more weights are given to them in the analysis [4,11]. Therefore, it is significant to evaluate approaches resolving the dependency issue in conducting meta-analyses.

Several methods were suggested to resolve the dependency issues. Multivariate meta-analysis was suggested as the most accurate procedure in handling dependency [12]. However, this method requires the correlation of the dependent effect sizes to be known to model the covariance matrix. Since this information is rarely reported by primary studies, implementing the multivariate meta-analysis is not feasible in practice. Performing separate meta-analyses, also known as shifting-unit-of-analysis, is believed to be another alternative to handling the dependency issue [13,14]. Instead of conducting one meta-analysis, this procedure accounts for the dependence by performing separate meta-analyses by multiple measures, multiple time-points, or multiple comparisons [1]. However, performing separate meta-analyses is not feasible in implementation unless measures or comparisons are consistent across primary studies. Choosing one effect size per study and then conducting a univariate meta-analysis can be another approach in handling the dependency issue [15]. The selection of the effect size can be based on one of the following rationales: (a) select one effect size randomly; (b) select the largest effect size; (c) select the effect size that is aligned with the focus of a meta-analysis; (d) select the effect size calculated by the measure with better validity and reliability statistics [4,16]. The procedure has appreciable limitations in that it reduces the statistical power by removing other effect sizes in one study [17]. In addition, it restricts the opportunity to examine the effect of within-study moderators on the variances of effect sizes.

The methods that are popular in practice and feasible in implementation include: averaging effect sizes within studies [1], using robust variance estimation (RVE) [18], and conducting a three-level meta-analysis (3LM) [11,17]. There were studies evaluating the performance of these methods with simulated effect sizes [2]. However, our knowledge of the efficacy of the three methods with real data is limited. Moreover, most studies dealt with effect sizes of standardized mean differences (i.e., Cohen's d or Hedge's g) extracted from experimental or quasi-experimental studies, which are different from correlational effect sizes in that the former refers to the treatment effects such as gains or differences between treatment and control groups whereas the latter refers to the magnitude of the relationships between two or more constructs. So far, no existing studies has examined different methods in handling dependent effect sizes based on correlation. To bridge the gap in the literature, the current study aims to compare the methods of the averaging method, 3LM, and working models with RVE, in dealing with dependent effect sizes of correlation with real data.

2. The Three Approaches

This section provides a review of the three approaches to synthesize dependent effect sizes: averaging within-study effect sizes, 3LM, and working models with RVE.

2.1. Averaging Within-Study Effect Sizes

The most straightforward approach to dealing with dependency is to compute average or weighted average effect sizes within studies before conducting a univariate meta-analysis. This method proceeds on the assumption that observed effect sizes in one study share the common effect in the population, which is reasonable but not the case in the strict sense. Due to its ease in implementation and conceptual understanding, the averaging method was commonly used in meta-analytic studies. In Ahn et al.'s [3] review of meta-analyses in education, 18 out of the 28 meta-analyses adopted the averaging method to synthesize dependent effect sizes.

The combined effect and the variance of the combined effect in one study can be expressed as follows.

$$\bar{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i, \tag{1}$$

$$V_{\bar{Y}} = \left(\frac{1}{m}\right)^2 \left(\sum_{i=1}^{m} V_i + \sum_{i \neq j} \left(r_{ij}\sqrt{V_i}\sqrt{V_j}\right)\right) \tag{2}$$

where Y_i is the i_{th} effect size of the m dependent effect sizes in one study, $\bar{Y}$ is the average effect size of the study, V_i and V_j are the variance for Y_i and Y_j, $V_{\bar{Y}}$ is the variance of the combined effect, and r_{ij} is the correlation between the two effect sizes, Y_i and Y_j. Notations are borrowed from Borenstein et al. [1].

According to the two formulas, while it is possible to calculate the mean effect sizes within studies, the variance of combined effects requires the correlations of the dependent effect sizes, which are not always reported in primary studies. There are several alternatives recommended by Borenstein et al. [1] when the correlations between dependent effect sizes are unknown: (a) identifying correlations based on empirical evidence or theoretical rationale; (b) assuming their correlation to be zero and treating them independently; (c) assuming their correlation to be one; (d) using the average of sampling variance of each effect size. Assuming their correlation to be zero would result in spuriously smaller variance, narrower confidence intervals, and a higher probability of committing a Type I error [1,2]. Conversely, assuming their correlation to be one has the risk of overestimating the variance of combined effect sizes, obtaining larger confidence intervals, and having a higher probability of committing a Type II error [1,4]. Most meta-analysts use the average of variances of dependent effect sizes to estimate the variance of the combined effect [18]. Therefore, the averaging procedure is reliable only when the dependent effect sizes are highly correlated. However, if dependent effect sizes are not homogenous within studies and the goal of meta-analysts is to detect this heterogeneity, the averaging procedure would obscure this rich information of within-study covariates [4,19].

2.2. Three-Level Meta-Analysis

Multilevel model is an optimal method in dealing with nested data since it can account for dependency and variances in multiple levels (e.g., student level, class level, and school level). This method has an application in meta-analytic data, which has the same hierarchical structure [20]. Literature suggested 3LM as an alternative to handling dependent effect sizes [11]. When the dependency comes from multiple outcomes, the first level is the participants' level, which models sampling variation for effect sizes for one outcome. The second level is the outcome level, varying within the outcome-specific errors. The third level is the study level, varying within the study-specific errors [17,21]. The three-level meta-analysis can be modeled as

$$\text{Level 1: } T_{iok} = \beta_{ok} + \varepsilon_{iok,} \tag{3}$$

$$\text{Level 2: } \beta_{ok} = \theta_k + \phi_{ok,} \tag{4}$$

$$\text{Level 3: } \theta_k = \gamma + \eta_{k,} \tag{5}$$

where T_{iok} is the estimate of effect size i for outcome o in study k, β_{ok} is the population parameter of effect size for outcome o in study k, ε_{iok} is the sampling error in level 1 with the mean of zero and variance of σ^2_{iok} (i.e., sampling variance), θ_k is the parameter for effect size in study k, and ϕ_{ok} is residual in outcome level with the mean of zero and variance of $\sigma^2_{\phi_{ok}}$ (i.e., between-outcome variance), γ is the parameter for the average effect size in population, η_k is the study-level residual with the mean of zero and the variance of $\sigma^2_{\eta_k}$ (i.e., between-study variance).

Fernández-Castilla et al. [22] systematically reviewed 178 meta-analyses published from 2002 to 2018 that used multilevel models and found that the majority of them employed a three-level model. Geeraert et al. [23] examined the effectiveness of early prevention programs on child abuse and neglect with 40 evaluation studies and performed a 3LM to synthesize dependent effect sizes from multiple outcomes within studies. The three levels account for the covariance of dependent effect sizes within studies by decomposing the total variance into between-effect variance and between-study variance. The 3LM has the advantage of exploring moderators for each level and the flexibility of the number of levels added to the hierarchical structure [2,11]. However, this method is better in handling hierarchical dependency than correlated dependency among effect sizes [18].

2.3. Working Models with Robust Variance Estimation

Accurate estimation of overall effect size necessitates covariances between dependent effect sizes, and this information is not always available in primary studies. Hedges et al. [18] solved the problem by proposing the RVE method. RVE does not require the underlying covariances to be known to calculate the variance of the average effect size for one study. Instead, it obtains a crude estimate of the covariances using the cross-product of residuals within one study. The robust variance estimator can be expressed as

$$V^R = \left(\sum_{j=1}^{m} X_j' W_j X_j\right)^{-1}\left(\sum_{j=1}^{m} X_j' W_j e_j e_j' W_j X_j\right)\left(\sum_{j=1}^{m} X_j' W_j X_j\right)^{-1}, \quad (6)$$

where e_j is the estimated residual vector for study *j*, m is the number of studies, W_j is the weight for study *j*. Hedges et al. [18] found that when the number of studies goes infinite, $e_j e_j'$ is a good estimator of the true covariance matrix. The most efficient weights and the hierarchical effects weight can be expressed as

$$\frac{1}{(V_{\cdot j}+\tau^2)(1+(k_j-1)\rho)}, \quad (7)$$

$$\frac{1}{k_j(\bar{V}_{\cdot j}+\hat{\tau}^2)} \quad (8)$$

where $V_{\cdot j}$ is the average of the variance of the effect size estimates in in study *j*, τ^2 is the between variance component, and ρ is the correlation between dependent effect sizes. Although ρ is unknown, a sensitivity approach can be taken to see if the results are affected by the choice of different ρ values (between 0 and 1).

The RVE approach has the following advantages compared to other methods in handling the dependency issue. First, RVE is advantageous to 3LM in that it is applicable to both hierarchical dependency and correlated dependency [6]. Second, it does not require the assumption of a normal distribution of effect sizes [2]. Third, RVE results are invariant regardless of the choice of weights or meta-analysis models [18]. Although RVE has the advantage of accurately estimating mean effects and variance without knowing correlations between dependent effect sizes within one study, it has stringent assumptions for accurate estimation. The implementation of RVE has a requirement on the minimum number of primary studies in a meta-analysis [6]. It performs well in estimating the mean effect size when there are ten or more studies involved [24]. It also requires that the difference between the number of primary studies and the number of covariates be more than four [25].

2.4. Literature in Comparing Different Approaches

Moeyaert et al. [2] compared three procedures in handling multiple outcomes: averaging effect sizes, RVE, and multilevel meta-analysis (MLM) with 432,000 simulated effect sizes under 216 conditions. The indices of overall effect estimates, standard error estimates, and variance estimates were used to evaluate the three procedures. The study found that while the three procedures functioned equally well in the estimation of overall fixed effects, RVE yielded unbiased estimation of variance under all conditions, whereas MLM underestimated the variance slightly when the number of primary studies was small (e.g., 25 studies). In the estimation of unbiased standard errors, both RVE and MLM were recommended.

Scammacca et al. [4] examined various procedures to handle dependency resulting from multiple measures and multiple group comparisons with real data in a meta-analysis of the effect of reading interventions on students with learning difficulties. Their procedures included choosing the highest effect size, randomly sampling an effect size, choosing a research-question-guided effect size, conducting meta-analyses separately, treating effect sizes independently, averaging effect sizes, using RVE, and employing a 3LM. They recommended the averaging method, RVE, and 3LM procedures when the overall effect size was aligned with the research purpose of a meta-analytic study. In addition, RVE and 3LM were preferred when the number of primary studies was large.

Most of the methodological studies examining methods dealing with dependent effect sizes used simulated data [2]. Although simulated data has the advantage of having a criterion against which different methods can be compared in terms of estimation accuracy, real data "can better emulate the types and nature of dependency that typically exist in studies that education researchers struggle to meta-analyze" [4] (p.336). In addition, previous studies with real data used the effect size of standardized mean difference, and their results may not be applied to effect sizes based on correlation [4]. Therefore, a methodological study is warranted to examine the three approaches in synthesizing dependent effect sizes based on correlation with real data.

The present study aims to compare three approaches to dealing with dependent effect sizes. Specifically, two research questions guided this study.

1. How do the averaging method, 3LM, and RVE differ in the estimation of overall mean, standard error, and heterogeneity of the true effect?
2. Do these estimates differ by the number of primary studies in a meta-analysis and the number of dependent effect sizes per study?

3. Methodology

3.1. Data

Data were from the meta-regression study examining the relationship between writing self-efficacy and writing achievement [26]. In the meta-analysis, primary studies were located using both electronic searches and manual searches. Seven electronic databases were targeted: Academic Search Complete, Education Research Complete, Eric, ProQuest Dissertations & Theses Global, PsycARTICLES, PsycINFO, and Web of Science. Studies were included if they were: (a) scholarly (peer-reviewed) articles or theses/dissertations; (b) published or written in the time frame of 1977–2019; (c) written in English; (d) examining students' writing self-efficacy; (e) quantitative empirical studies; (f) studies that reported or had computable effect sizes. Please refer to the study [26] for detailed information for literature searches, screening, and coding. This final sample consisted of 565 effect sizes nested within 76 primary studies.

3.2. Effect Size Calculation

The effect size used in the current study is Pearson correlation coefficient, *r*. Studies that reported standardized beta coefficient (β) were also included by converting β to *r* metric using the two-parameter least-square equation.

$$r = 0.98\ \beta + 0.05\ \lambda, \tag{9}$$

where λ is an indicator variable, the value of which is 1 when β is nonnegative and 0 when β is negative [27]. Then *r* was transformed to Fisher's z metric using the formula $z = 0.5 \times \ln(\frac{1+r}{1-r})$ for the normalization of the effect size distribution [1]. The final sample consisted of 76 primary studies and 565 effect sizes, with an average of 7 effect sizes per study. The frequency of the number of effect sizes per study was presented in Table 1.

Table 1. Frequency of the Number of Effect Sizes per Study.

# of Effect Sizes	Frequency	Percent
1	15	0.20
2	13	0.17
3	6	0.08
4	12	0.16
5	4	0.05
6	11	0.14
9	1	0.01
10	2	0.03
11	1	0.01
12	2	0.03
13	1	0.01
14	1	0.01
18	2	0.03
27	1	0.01
28	1	0.01
42	1	0.01
56	1	0.01
92	1	0.01

3.3. Sensitivity Analysis

To generate more data sets for sensitivity analysis, various conditions were created to examine if the performance of the three methods differs by the number of primary studies and the number of effect sizes per study. Three levels of the number of effect sizes per study (*K*) were manipulated ($K \geq 2$, $K \geq 4$, and $K \geq 6$), which yielded three levels of the number of primary studies (*N*) of 61, 42, and 26. The three values can be used to represent large-, medium-, and small-scale meta-analyses, respectively. Based on the three levels of *K* and the three levels of *N*, six datasets were created for analysis. The design of the conditions was presented in Table 2, where 61-2 represents the condition of 61 primary studies with 2 effect sizes or more for each study. For the

conditions 42-2, 26-2, and 26-4, stratified random sampling was performed to select the sampled studies. Stratification was based on the number of effect sizes per study. The condition of all effect sizes was also used, so there were seven conditions in total.

Table 2. Six Conditions by Number of Primary Studies and Number of Effect Sizes per Study.

	# Studies 61	**# Studies 42**	**# Studies 26**
# of ES $\geq$ 2	Condition 61-2	Condition 42-2	Condition 26-2
# of ES $\geq$ 4	NA	Condition 42-4	Condition 26-4
# of ES $\geq$ 6	NA	NA	Condition 26-6

Note. ES = Effect Size. NA = not applicable, for example, there are only 26 studies when the number of effect sizes per study is equal to or greater than 6.

3.4. Data Analysis

A random-effects model was chosen for analysis because the assumption that true effects may vary across studies is more plausible. A univariate meta-analysis (UVM, assuming all the effect sizes independent) was performed as the baseline methods against which the three methods were compared. Each of the four methods (i.e., the univariate method, the averaging method, 3LM, and working model with RVE) was applied to the seven datasets, respectively. In the averaging method, the variance of the combined effect size in each study was calculated by averaging the variances of dependent effect sizes. For RVE, five levels of correlations of within-study dependent effect sizes (i.e., $\rho = 0$, $\rho = 0.1$, $\rho = 0.3$, $\rho = 0.5$, and $\rho = 1$) were employed to see if the results were robust to different values of ρ. Analyses were conducted using computer software R with different packages. The univariate and the averaging method were applied using R package metafor [28]. 3LM and correlated effects working model with RVE were implemented by using R packages metaSEM [16] and robumeta [29], respectively.

3.5. Evaluation Indices

Evaluation indices include the mean effect size, *Z* and its *p*-value, the standard error of the mean effect size, confidence intervals, and three heterogeneity statistics (Q, T^2, and I^2). Z and its *p*-values test the statistical significance of the mean effect size (i.e., whether it is statistically significantly different from zero). The statistics of standard error and confidence intervals measure the precision of the effect size estimate. Heterogeneity statistics quantify the extent to which true effect size varies across studies. Q statistic and its *p*-value test the statistical significance of heterogeneity (i.e., whether the true effect sizes vary across studies). Q can be obtained by computing the weighted sum of squares of the deviation of effect size in each study from the overall average effect size. T^2 is the estimate of the magnitude of the variance of the true effect sizes across studies (τ^2). I^2 refers to the proportion of variation due to heterogeneity, which can be calculated by dividing the between-study variance by the total variance. The present study takes Higgins et al.'s benchmark, and I^2 was interpreted as small when $I^2 =$ 25%, moderate when $I^2 =$ 50%, and high when $I^2 =$ 75% [30].

4. Results

4.1. Condition of All Effect Sizes

The effect size and heterogeneity statistics for all the effect sizes were presented in Table 3. The overall effect size and standard error estimates generated by the averaging method and RVE were very similar. 3LM yielded a slightly smaller effect size and standard error estimates, whereas the univariate method produced a much smaller effect size (0.2631) and standard error (0.0087) compared with the other four methods. As for the heterogeneity statistic, all the methods had statistically significant Q-values, indicating that the true effect sizes were heterogeneous across studies regardless of methods employed. All the I^2 statistics suggested that the majority of observed variance was due to between-study variation. 3LM yielded the largest estimate of true effect sizes variance ($T^2 = 0.0596$), and the univariate method yielded the smallest ($T^2 = 0.0347$). In addition, the discrepancy in the variance estimates between the averaging method ($T^2 = 0.0579$), 3LM method, and RVE ($T^2 = 0.0518$) were negligible. The sensitivity test for RVE indicated that the results did not vary based on the choice of ρ value.

Table 3. Effect Size and Heterogeneity Statistics for the Condition of All the Effect Sizes.

	Effect size and 95% Confidence Interval								Heterogeneity	
	k	ES (z)	ES (r)	SE	95% CI	Z	p	p of Q	I^2	T^2
UVM	565	0.2631	0.2572	0.0087	[0.2460, 0.2802]	30.1467	<0.001	<0.001	89.34%	0.0347
AVM	76	0.3179	0.3076	0.0297	[0.2598, 0.3760]	10.7208	<0.001	<0.001	92.36%	0.0579
3LM	76	0.3128	0.3030	0.0267	[0.2604, 0.3652]	11.701	<0.001	<0.001	21.70%/71.82%	0.0138/0.0458
ρ = 0	76	0.3183	0.3080	0.0299	[0.2587, 0.3778]	10.6511	<0.001	<0.001	91.08%	0.0518
ρ = 0.1	76	0.3183	0.3080	0.0299	[0.2587, 0.3778]	10.6511	<0.001	<0.001	91.08%	0.0518
RVE ρ = 0.3	76	0.3183	0.3080	0.0299	[0.2587, 0.3778]	10.6511	<0.001	<0.001	91.10%	0.0518
ρ = 0.5	76	0.3183	0.3080	0.0299	[0.2587, 0.3778]	10.6511	<0.001	<0.001	91.12%	0.0518
ρ = 1	76	0.3183	0.3080	0.0299	[0.2587, 0.3778]	10.6511	<0.001	<0.001	91.16%	0.0518

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.2. Condition of 61-2

See Table 4 for the results for the condition of having 61 primary studies with at least 2 effect sizes in each study. RVE generated the highest value of overall effect size (0.03017) and standard error (0.0254), followed by the averaging method and the 3LM. Overall, the three methods produced very similar values in the two statistics. Similar to the condition of all effect sizes, the univariate method had the smallest estimate of overall average effect size (0.2600) and standard error (0.0085). 3LM (T^2 = 0.0430) and RVE (T^2 = 0.0399) yielded the largest true variance estimates across studies, whereas the averaging method produced the smallest (T^2 = 0.0299). RVE results were robust to the choice of ρ value.

Table 4. Effect Size and Heterogeneity Statistics for the Condition of 61-2.

	Effect Size and 95% Confidence Interval								Heterogeneity	
	k	ES (z)	ES (r)	SE	95% CI	Z	p	p of Q	I^2	T^2
UVM	550	0.2600	0.2543	0.0085	[0.2434, 0.2765]	30.7570	<0.001	<0.001	88.38%	0.0311
AVM	61	0.3008	0.2920	0.0249	[0.2520, 0.3496]	12.0785	<0.001	<0.001	87.06%	0.0299
3LM	61	0.2983	0.2898	0.0238	[0.2516, 0.3450]	12.5199	<0.001	<0.001	29.11%/62.19%	0.0137/0.0293
ρ = 0	61	0.3017	0.2929	0.0254	[0.2508, 0.3527]	11.8605	<0.001	<0.001	89.37%	0.0399
ρ = 0.1	61	0.3017	0.2929	0.0254	[0.2508, 0.3527]	11.8605	<0.001	<0.001	89.38%	0.0399
RVE ρ = 0.3	61	0.3017	0.2929	0.0254	[0.2508, 0.3527]	11.8605	<0.001	<0.001	89.41%	0.0399
ρ = 0.5	61	0.3017	0.2929	0.0254	[0.2508, 0.3527]	11.8605	<0.001	<0.001	89.44%	0.0400
ρ = 1	61	0.3017	0.2929	0.0254	[0.2508, 0.3527]	11.8605	<0.001	<0.001	89.51%	0.0400

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.3. Condition of 42-2

The effect size and heterogeneity statistics for the condition of having 42 primary studies with at least 2 effect sizes per study were shown in Table 5. Like the condition of 61-2, the averaging method, 3LM, and RVE produced very similar combined effect size estimates and standard error estimates, with RVE having a slightly larger value than the other two methods. The univariate method was the lowest in the estimates of the combined effect size (0.2738) and standard error (0.0108). The univariate method (T^2 = 0.341) resulted in a similar true variance estimate as was obtained by RVE (T^2 = 0.331) or the 3LM (T^2 = 0.389) procedures. The averaging method resulted in the smallest true variance estimate (T^2 = 0.0221).

Table 5. Effect Size and Heterogeneity Statistics for the Condition of 42-2.

	Effect size and 95% Confidence Interval								Heterogeneity	
	k	ES (z)	ES (r)	SE	95% CI	Z	p	p of Q	I^2	T^2
UVM	359	0.2738	0.2672	0.0108	[0.2527, 0.2950]	25.4016	<0.001	<0.001	90.28%	0.0341
AVM	42	0.3096	0.3001	0.0266	[0.2575, 0.3618]	11.6383	<0.001	<0.001	82.86%	0.0221
3LM	42	0.3064	0.2972	0.0263	[0.2549, 0.3578]	11.6658	<0.001	<0.001	36.40%/54.95%	0.0155/0.0234
ρ = 0	42	0.3109	0.3013	0.0271	[0.2562, 0.3657]	11.4836	<0.001	<0.001	86.92%	0.0331
ρ = 0.1	42	0.3109	0.3013	0.0271	[0.2562, 0.3657]	11.4836	<0.001	<0.001	86.94%	0.0331
RVE ρ = 0.3	42	0.3109	0.3013	0.0271	[0.2562, 0.3657]	11.4836	<0.001	<0.001	86.99%	0.0331
ρ = 0.5	42	0.3109	0.3013	0.0271	[0.2562, 0.3657]	11.4836	<0.001	<0.001	87.04%	0.0331
ρ = 1	42	0.3109	0.3013	0.0271	[0.2562, 0.3657]	11.4836	<0.001	<0.001	87.17%	0.0331

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.4. Condition of 26-2

The results for the condition of having 26 primary studies with at least 2 effect sizes per study were presented in Table 6. RVE resulted in the highest values of effect size estimate (0.3175) and standard error (0.0341), followed by the averaging method. The univariate method produced the smallest value in both the overall effect size (0.2924) and standard error estimate (0.0158). All the overall effect sizes were statistically significantly different from zero, provided evidence of the statistical significance of the effect size. The heterogeneity statistics showed that the 3LM resulted in the highest value of the true variance estimate (T^2 = 0.0363), followed by RVE (T^2 = 0.0326). The averaging method produced a much smaller variance estimate (T^2 = 0.0227) compared to the other three methods.

Table 6. Effect Size and Heterogeneity Statistics for the Condition of 26-2.

		Effect Size and 95% Confidence Interval							Heterogeneity	
	k	**ES (*z*)**	**ES (*r*)**	**SE**	**95% CI**	***Z***	***p***	***p* of Q**	I^2	T^2
UVM	160	0.2924	0.2843	0.0158	[0.2615, 0.3234]	18.5056	<0.001	<0.001	91.44%	0.0334
AVM	26	0.3154	0.3053	0.0336	[0.2495, 0.3813]	9.3845	<0.001	<0.001	84.72%	0.0227
3LM	26	0.3124	0.3026	0.0331	[0.2475, 0.3773]	9.4401	<0.001	<0.001	32.18%/59.88%	0.0127/0.0236
ρ = 0	26	0.3175	0.3072	0.0341	[0.2472, 0.3878]	9.3203	<0.001	<0.001	87.90%	0.0326
ρ = 0.1	26	0.3175	0.3072	0.0341	[0.2472, 0.3878]	9.3203	<0.001	<0.001	87.93%	0.0327
RVE ρ = 0.3	26	0.3175	0.3072	0.0341	[0.2472, 0.3878]	9.3203	<0.001	<0.001	88.01%	0.0327
ρ = 0.5	26	0.3175	0.3072	0.0341	[0.2472, 0.3878]	9.3203	<0.001	<0.001	88.09%	0.0327
ρ = 1	26	0.3175	0.3072	0.0341	[0.2472, 0.3878]	9.3203	<0.001	<0.001	88.27%	0.0328

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.5. Condition of 42-4

For the condition of having 42 primary studies with at least 4 effect sizes per study (see Table 7), the averaging method had the largest effect size estimate (0.2747), whereas the RVE had the largest standard error estimate (0.0258). In general, the averaging method, 3LM and RVE estimated somewhat similar values of the overall effect size and standard error. The univariate method estimated the smallest values of the overall effect size and standard error. As for the variance of true effect sizes across studies, the 3LM (T^2 = 0.0343) and RVE (T^2 = 0.0337) produced results with negligible differences. However, the averaging method had a much smaller variance estimate (T^2 = 0.0194). RVE results were also robust to the choice of ρ value.

Table 7. Effect Size and Heterogeneity Statistics for the Condition of 42-4.

		Effect Size and 95% Confidence Interval							Heterogeneity	
	k	**ES (*z*)**	**ES (*r*)**	**SE**	**95%CI**	***Z***	***p***	***p* of Q**	I^2	T^2
UVM	506	0.2516	0.2464	0.0083	[0.2352, 0.2679]	30.2345	<0.001	<0.001	86.99%	0.0271
AVM	42	0.2747	0.2680	0.0253	[0.2252, 0.3242]	10.8762	<0.001	<0.001	82.06%	0.0194
3LM	42	0.2738	0.2672	0.0243	[0.2260, 0.3216]	11.2294	<0.001	<0.001	33.09%/56.36%	0.0127/0.0216
ρ = 0	42	0.2744	0.2677	0.0258	[0.2223, 0.3265]	10.6	<0.001	<0.001	87.90%	0.0337
ρ = 0.1	42	0.2744	0.2677	0.0258	[0.2223, 0.3265]	10.6	<0.001	<0.001	87.93%	0.0337
RVE ρ = 0.3	42	0.2744	0.2677	0.0258	[0.2223, 0.3265]	10.6	<0.001	<0.001	87.98%	0.0338
ρ = 0.5	42	0.2744	0.2677	0.0258	[0.2223, 0.3265]	10.6	<0.001	<0.001	88.03%	0.0338
ρ = 1	42	0.2744	0.2677	0.0258	[0.2223, 0.3265]	10.6	<0.001	<0.001	88.15%	0.0338

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.6. Condition of 26-4

The results of effect size and heterogeneity estimates for the condition having 26 primary studies with at least 4 effect sizes per study were shown in Table 8. All the methods resulted in statistically significant effect sizes ($p < 0.001$), with the 3LM having the largest (0.2939) and the univariate method having the smallest value (0.2743). Like the aforementioned conditions, the averaging method, 3LM, and RVE yielded somewhat similar overall effect size and standard error estimates. As for the heterogeneity statistics, all the methods generated statistically significant Q-values, indicating that the true effects varied across studies, and more than 60% of the observed variance was between-study variance (I^2 ranging from 62.90% to 93.37%). The averaging method had the smallest variance estimate (0.0264) as compared with the other four methods.

Table 8. Effect Size and Heterogeneity Statistics for the Condition of 26-4.

	Effect Size and 95% Confidence Interval							Heterogeneity		
	k	**ES (*z*)**	**ES (*r*)**	**SE**	**95%CI**	***Z***	***p***	***p* of Q**	I^2	T^2
UVM	180	0.2743	0.2676	0.0156	[0.2438, 0.3048]	17.6114	<0.001	<0.001	92.45%	0.0374
AVM	26	0.2934	0.2853	0.0359	[0.2229, 0.3638]	8.1656	<0.001	<0.001	87.02%	0.0264
3LM	26	0.2918	0.2838	0.0357	[0.2218, 0.3618]	8.1719	<0.001	<0.001	29.70%/63.67%	0.0137/0.0293
ρ = 0	26	0.2939	0.2857	0.0364	[0.2189, 0.3689]	8.0799	<0.001	<0.001	89.87%	0.0380
ρ = 0.1	26	0.2939	0.2857	0.0364	[0.2189, 0.3689]	8.0799	<0.001	<0.001	89.90%	0.0381
RVE ρ = 0.3	26	0.2939	0.2857	0.0364	[0.2189, 0.3689]	8.0799	<0.001	<0.001	89.97%	0.0381
ρ = 0.5	26	0.2939	0.2857	0.0364	[0.2189, 0.3689]	8.0799	<0.001	<0.001	90.38%	0.0381
ρ = 1	26	0.2939	0.2857	0.0364	[0.2189, 0.3689]	8.0799	<0.001	<0.001	90.21%	0.0382

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

4.7. Condition of 26-6

Table 9 presents the results for the condition of having 26 primary studies with at least 6 effect sizes per study. All the Z values were statistically significant ($p < 0.001$), suggesting that the effect size estimates were statistically significantly different from zero, regardless of the methods employed. There were small differences in the estimates of overall effect size and standard error between the averaging method, 3LM, and RVE, whereas the univariate method generated much smaller values in the two estimates. Q-values indicated that the true effect sizes varied across studies. RVE resulted in the largest variance (0.0350), which was similar to the results produced by 3LM (0.0301). However, the averaging method produced a much smaller estimate of the variance of true effect sizes across studies (0.0164). The sensitivity test for RVE indicated that the results did not vary by the ρ value

Table 9. Effect Size and Heterogeneity Statistics for the Condition of 26-6.

	Effect size and 95% confidence interval							Heterogeneity		
	k	**ES (*z*)**	**ES (*r*)**	**SE**	**95%CI**	**Z**	***p***	***p of Q***	I^2	T^2
UVM	438	0.2484	0.2434	0.0088	[0.2313, 0.2656]	28.3838	<0.001	<0.001	86.97%	0.0258
AVM	26	0.2799	0.2728	0.0297	[0.2217, 0.3382]	9.4228	<0.001	<0.001	81.83%	0.0164
3LM	26	0.2734	0.2668	0.0271	[0.2203, 0.3265]	10.0945	<0.001	<0.001	39.23%/40.39%	0.0133/0.0168
ρ = 0	26	0.2769	0.2700	0.0300	[0.2150, 0.3387]	9.2349	<0.001	<0.001	89.72%	0.0350
ρ = 0.1	26	0.2769	0.2700	0.0300	[0.2150, 0.3387]	9.2349	<0.001	<0.001	89.46%	0.0350
RVE ρ = 0.3	26	0.2769	0.2700	0.0300	[0.2150, 0.3387]	9.2349	<0.001	<0.001	89.50%	0.0351
ρ = 0.5	26	0.2769	0.2700	0.0300	[0.2150, 0.3387]	9.2349	<0.001	<0.001	89.60%	0.0351
ρ = 1	26	0.2769	0.2700	0.0300	[0.2150, 0.3387]	9.2349	<0.001	<0.001	89.79%	0.0352

Note. ρ is the correlation of within-study effect size; ES =effect size; SE =standard error; CI = confidence intervals; UNM = univariate method; AVM = averaging method; 3LM = three-level meta-analysis; RVE = robust variance estimation.

5. Discussion

5.1. Similarities Between the Three Approaches

There are three similarities in the estimation results between the five methods across all the conditions. First, all the methods resulted in statistically significant *Z*-values, providing evidence of the statistical significance of the overall effect size. This finding suggested that all the three approaches were legitimate in meta-analyses. Second, all the methods produced significant Q-values across different conditions, providing evidence that the true effect sizes were heterogeneous across studies. This confirmed the legitimacy of the use of a random-effects model. The significant Q-values suggest the untenability of the null hypothesis that all the studies share the common true effect size or a small amount of observed dispersion of effect sizes with precise studies [1]. Therefore, this finding can be interpreted as either a large amount of dispersion of the true effect sizes across studies or a little amount of dispersion with a precise estimation of effect sizes. When the effect size is Fisher's z, the precision of estimation depends on the inverse of the sample size ($\frac{1}{N-3}$) [1]. The current study has a mean sample size of 245.03 (SD = 247.77). The large sample size may account for the significant Q-values estimated. Third, all three approaches produced large I^2 statistics (ranging from 79.62% to 92.45%). According to Higgins et al.'s benchmark [30], it can be interpreted as a large value (if larger than 75%), suggesting that a substantial proportion of the observed variation reflected real heterogeneity or the variation of true effect sizes across studies. The large between-study variance suggested the necessity of conducting moderator analyses in the future to examine what factors at the study level may account for the heterogeneity.

5.2. Differences Between the Three Approaches

5.2.1. Overall Average Effect Size

RVE resulted in the highest values of overall average effect size estimates, followed by the averaging method. This pattern was consistent across the seven conditions except for the conditions of 42-4 and 26-6, in which the averaging method had slightly larger effect size estimates than those estimated by RVE. Overall, there were negligible differences in the overall effect size estimates between the averaging method, 3LM, and RVE. However, the univariate method resulted in much smaller effects as compared with the other three methods across the seven conditions. This result was consistent with Moeyaert et al.'s study [2], which concluded that the averaging method, RVE, and MLM all produced unbiased effect size estimates. The researchers further noted that this was not impacted by the number of primary studies or the number of effect sizes per study. This finding was also partially consistent with Scammacca's study [4], which found that RVE and the averaging method produced similar effect size estimates. Moeyaert et al. and Scammacca's studies examined the effect sizes of standardized mean differences [2,4]. The current study extends their findings to the effect size of correlation (Fisher' z).

5.2.2. Standard Error

There were little differences between the averaging method, 3LM, and RVE in terms of standard error estimates, although RVE resulted in slightly larger values while the averaging method resulted in slightly smaller values. The univariate method produced much smaller results of standard errors. The performance of the three approaches was consistent across the seven conditions. Moeyaert et al. found that RVE, MLM, and the averaging method produced unbiased standard errors in most conditions except for the condition of the number of effect sizes of 4 and the correlation between within-study effect sizes of 0, in which the averaging method overestimated standard errors by 36% [2]. The discrepancy between the current study and Moeyaert et al.'s study may be due to the use of real data in the current study, where within-study effect sizes are correlated to some extent. The small value of estimated standard error yielded by the univariate method is expected. If we treat dependent effect sizes as independent and analyze them one by one, the variance of the overall effect size will be underestimated. This would result in a narrow confidence interval for the combined effect size and increase the likelihood of a Type I error.

5.2.3. Variance of True Effect Size Estimates

3LM yielded the largest variance estimates, followed by the RVE, whereas the averaging method estimated a much smaller variance. The value produced by the 3LM was similar to the value produced by RVE. This pattern was consistent across the conditions except for the condition of all the effect sizes and the condition of 26-6. This finding was partially consistent with Moeyaert et al.'s study [2], which noted that the averaging method extremely underestimated the variance estimate. Given that the averaging method was implemented by averaging the within-study dependent effect sizes, the variance of these within-study effect sizes was removed in this process. On the contrary, the RVE estimated the total variances, and 3LM estimated the two variances (i.e., between-study variance and between-outcome variance) separately. The current study also found that the variance estimated by 3LM was slightly smaller than RVE for the condition of 26-6. This was also similar to Moeyaert et al.'s finding that 3LM slightly underestimated variances when the number of studies was less than 25 [2].

6. Limitations and Future Research

The limitation of the present study is the use of real data only, which limited the opportunity of having a criterion against which to evaluate the estimation accuracy of the three approaches. Future studies are recommended to use simulated data to investigate which method results in unbiased estimates of effect size and heterogeneity statistics. Second, the current study did not examine the effect of sources of dependency (e.g., multiple outcomes, multiple time-points) on the efficacy of the three approaches. Dependent effect sizes from various outcome measures differ from those being measured at multiple time-points in data structure and correlation matrices, which may interact with the performance of these methods. Future studies are suggested to compare the methods in dealing with a certain type of dependency.

7. Implications and Conclusions

We found that the true effect sizes were heterogeneous across studies, and a substantial amount of variance was between-study variance regardless of methods employed. This finding suggested the necessity of conducting moderator analyses to explore what factors may account for heterogeneity.

The study has implications for meta-analysts. The current study found the univariate method produced much smaller estimates of the standard error when synthesizing dependent effect sizes, which could result in spuriously statistically significant effect sizes. Therefore, meta-analysts need to report whether the effect sizes are independent or not and avoid using the univariate method if the dependency issue occurs. However, this information is not always available, and the dependency issue is routinely ignored by meta-analysts [3,31]. In Ahn's review of meta-analyses in education, 27% of meta-analyses did not provide information about whether the effect sizes are independent [3]. The current study also provides meta-analysts with insights into choosing appropriate methods in addressing the dependency issue. While the three methods, the averaging method, 3LM, and RVE had similar results of overall effect size and standard error estimates, they performed differently in estimating variances. The averaging method had much smaller variances and the 3LM yielded the smallest value of variance for the condition of 26-6. A falsely small variance would conceal potential factors that may explain the heterogeneity. Therefore, meta-analysts are not recommended to adopt the averaging method, and the 3LM is not preferred when the number of studies is small.

The study also has implications for primary researchers. First, primary researchers can assist meta-analysts by providing effect size information in addition to the information of statistical significance. The recommendation of reporting effect size information can be found in Educational and Psychological Measurement [32] and a series of versions (4th, 5th, 6th, 7th) of American Psychological Association Publication Manual [33–36]. Lack of the effect size information or sufficient statistics to compute effect sizes would invalidate the eligibility of a primary study in a meta-analysis and decrease the statistical power of the meta-analysis. Second, when a primary study has multiple outcome measures or multiple comparisons, primary researchers need to provide detailed information on these measures or treatments. Meta-analysts will benefit by knowing how these measures are conceptualized and operationalized as well as how these treatments are implemented. The extent to which these measures or treatments are different helps meta-analysts to choose appropriate approaches to addressing the dependency issue. Primary researchers are also recommended to provide correlations between multiple measures, which is necessary for modeling the dependency

Author Contributions

T.S.: conceptualization, methodology, software, data curation, writing-original draft preparation, visualization, investigation, writing-reviewing and editing; C.W.: conceptualization, methodology, writing-reviewing and editing, supervision, validation; R.G.L.: conceptualization, methodology, writing-reviewing and editing, supervision, validation. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author and can be shared upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Borenstein, M.; Hedges, L.V.; Higgins, J.P.; et al. *Introduction to Meta-Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
2. Moeyaert, M.; Ugille, M.; Beretvas, S.N.; et al. Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *Int. J. Soc. Res. Methodol.* **2017**, *20*, 559–572. https://doi.org/10.1080/13645579.2016.1252189.
3. Ahn, S.; Ames, A.J.; Myers, N.D. A review of meta-analyses in education: Methodological strengths and weaknesses. *Rev. Educ. Res.* **2012**, *82*, 436–476. http://doi.org/10.3102/0034654312458162.
4. Scammacca, N.; Roberts, G.; Stuebing, K.K. Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Rev. Educ. Res.* **2014**, *84*, 328–364. http://doi.org/10.3102/0034654313500826.
5. Cheung, S.F.; Chan, D. Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *J. Appl. Psychol.* **2004**, *89*, 780–791. https://doi.org/10.1037/0021-9010.89.5.780.
6. Tanner-Smith, E.; Tipton, E. Robust variance estimation with dependent effect sizes: Practical considerations and a software tutorial in Stata and SPSS. *Res. Synth. Methods* **2014**, *5*, 13–30. https://doi.org/10.1002/jrsm.1091.
7. Lambert, R.G. The ROBUSTNESS of the Standard Errors of Summarized, Corrected Validity Coefficients to Non-Independence and Non-Normality of Primary Data. Doctoral Dissertation, Georgia State University, Atlanta, GA, USA, 1995.
8. Hashemnejad, F.; Zoghi, M.; Amini, D. The relationship between self-efficacy and writing performance across genders. *Theory Pract. Lang. Stud.* **2014**, *4*, 1045–1052. https://doi.org/10.4304/tpls.4.5.1045-1052.
9. Perin, D.; Lauterbach, M.; Raufman, J.; et al. Text-based writing of low-skilled postsecondary students: Relation to comprehension, self-efficacy and teacher judgments. *Read. Writ.* **2017**, *30*, 887–915. http://doi.org/10.1007/s11145-016-9706-0.
10. Shintani, N.; Aubrey, S. The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *Mod. Lang. J.* **2016**, *100*, 296–319. http://doi.org/10.1111/modl.12317.
11. Van den Noortgate, W.; López-López, J.A.; Marín-Martínez, F.; et al. Three-level meta-analysis of dependent effect sizes. *Behav. Res. Methods* **2013**, *45*, 576–594. http://doi.org/10.3758/s13428-012-0261-6.
12. Kalaian, H.A.; Raudenbush, S.W. A multivariate mixed linear model for meta-analysis. *Psychol. Methods* **1996**, *1*, 227–235. http://doi.org/10.1037/1082-989X.1.3.227.
13. Cooper, H.M. *Synthesizing Research: A Guide for Literature Reviews*, 3rd ed.; Sage: Thousand Oaks, CA, USA, 1998.
14. Cooper, H.M. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 4th ed.; Sage: Thousand Oaks, CA, USA, 2010.
15. Chambers, E.A. An introduction to meta-analysis with articles from the Journal of Educational Research (1992–2002). *J. Educ. Res.* **2004**, *98*, 35–45. https://doi.org/10.3200/JOER.98.1.35-45.
16. Cheung, M.W.L. A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychol. Rev.* **2019**, *29*, 387–396. http://doi.org/10.1007/s11065-019-09415-6.
17. Cheung, M.W.L. Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychol. Methods* **2014**, *19*, 211–229. http://doi.org/10.1037/a0032968.
18. Hedges, L.V.; Tipton, E.; Johnson, M.C. Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Methods* **2010**, *1*, 39–65. http://doi.org/10.1002/jrsm.5.
19. Marín-Martínez, F.; Sánchez-Meca, J. Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *Span. J. Psychol.* **1999**, *2*, 32–38. http://doi.org/10.1017/S1138741600005436.
20. Raudenbush, S.W.; Bryk, A.S. *Hierarchical Linear Models: Applications and Data Analysis Methods*; Sage: Thousand Oaks, CA, USA, 2002.
21. Konstantopoulos, S. Fixed effects and variance components estimation in three-level meta-analysis. *Res. Synth. Methods* **2011**, *2*, 61–76. http://doi.org/10.1002/jrsm.35.
22. Fernández-Castilla, B.; Jamshidi Declercq, L.; Beretvas, S.N.; et al. The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behav. Res. Methods* **2020**, *52*, 2031–2052. https://doi.org/10.3758/s13428-020-01373-9.
23. Geeraert, L.; VandenNoortgate, W.; Grietens, H.; et al. The effects of early prevention programs for families with young children at risk for physical child abuse and neglect: A meta-analysis. *Child Maltreatment* **2004**, *9*, 277–291. https://doi.org/10.1177/1077559504264265.
24. Tipton, E. Robust variance estimation in meta-regression with binary dependent effects. *Res. Synth. Methods* **2013**, *4*, 169–187. https://doi.org/10.1002/jrsm.1070.
25. Tanner-Smith, E.; Tipton, E.; Polanin, J.R. Handling complex meta- analytic data structures using robust variance estimates: A tutorial in R. *J. Dev. Life-Course Criminol.* **2016**, *2*, 85–112. http://doi.org/10.1007/s40865-016-0026-5.

26. Sun, T.; Wang, C.; Lambert, R.G.; Liu, L. Relationship between second language English writing self-efficacy and achievement: A meta-regression analysis. *J. Second. Lang. Writ.* **2021**, *53*, 100817.
27. Peterson, R.A.; Brown, S.P. On the use of beta coefficients in meta-analysis. *J. Appl. Psychol.* **2005**, *90*, 175–181.
28. Viechtbauer, W. Conducting meta-analyses in R with the metaphor package. *J. Stat. Softw.* **2010**, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03.
29. Fisher, Z.; Tipton, E. Robumeta: Robust Variance Meta-Regression. 2014. Available online: http://cran.rproject.org/web/packages/robumeta/index.html (accessed on 1 September 2020).
30. Higgins, J.P.; Thompson, S.G.; Deeks, J.J.; et al. Measuring inconsistency in meta-analyses. *BMJ* **2003**, *327*, 557–560. http://doi.org/10.1136/bmj.327.7414.557.
31. Rodgers, M.A.; Pustejovsky, J.E. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychol. Methods* **2021**, *26*, 141.
32. Thompson, B. Guidelines for authors. *Educ. Psychol. Meas.* **1994**, *54*, 837–847.
33. American Psychological Association. *Publication Manual of the American Psychological Association*, 4th ed.; American Psychological Association: Washington, DC, USA, 1994.
34. American Psychological Association. *Publication Manual of the American Psychological Association*, 5th ed.; American Psychological Association: Washington, DC, USA, 2001.
35. American Psychological Association. *Publication Manual of the American Psychological Association*, 6th ed.; American Psychological Association: Washington, DC, USA, 2010.
36. American Psychological Association. *Publication Manual of the American Psychological Association*, 7th ed.; American Psychological Association: Washington, DC, USA, 2019.