Journal of Educational Psychology
2015, Vol. 107, No. 2, 558–579

0022-0663/15/$12.00 http://dx.doi.org/10.1037/a0037698

Preliminary Empirical Model of Crucial Determinants of Best Practice for Peer Tutoring on Academic Achievement

Kim Chau Leung
Hong Kong Institute of Education

Previous meta-analyses of the effects of peer tutoring on academic achievement have been plagued with theoretical and methodological flaws. Specifically, these studies have not adopted both fixed and mixed effects models for analyzing the effect size; they have not evaluated the moderating effect of some commonly used parameters, such as comparing same-age reciprocal peer tutoring, same-age nonreciprocal, or cross-age peer tutoring; considered the educational level of tutee or tutor; or properly addressed publication bias. Most studies are confined to specific populations and particular subjects (mainly mathematics and reading), and some studies are confounded by other types of intervention (such as cooperative learning or adult-led tutoring). Hence, there is a compelling need for an updated, comprehensive meta-analysis evaluating the effect of peer tutoring on academic achievement that incorporates advances in methodology, is not confounded by other modes of peer learning, and engages a wide range of participants and various subjects. The present study demonstrates that peer tutoring has a positive impact on academic achievement. The moderators and crucial determinants of the effectiveness of peer tutoring are identified and compared. Moreover, program parameters based on the concepts of role theory and interdependent group contingencies are evaluated. Finally, a preliminary empirical model of the crucial determinants of best practices for peer tutoring on academic achievement is proposed.

Keywords: meta-analysis, peer tutoring, tutee, achievement

Supplemental materials: http://dx.doi.org/10.1037/a0037698.supp

Peer tutoring has been commonly implemented in educational settings. Research has demonstrated that peer tutoring has a positive impact on academic outcomes. Several meta-analyses have examined the effects of peer tutoring systematically and empirically. However, many of these meta-analyses have been plagued with methodological or theoretical flaws. The aim in the present investigation was to synthesize previous research on peer tutoring, address its limitations, evaluate the effects of peer tutoring on academic achievement, and identify the crucial determinants of the effectiveness of peer tutoring. Finally, a preliminary empirical model of the crucial determinants of best practices for peer tutoring on academic achievement is proposed.

Importance of Meta-Analysis

Numerous reviews (e.g., Fuchs et al., 2008; McMaster, Fuchs, & Fuchs, 2006; Robinson, Schofield, & Steers-Wentzell, 2005; Scruggs & Ritcher, 1988; Spencer & Balboni, 2003; Topping, 1987, 1996, 2005) have demonstrated that peer tutoring has positive impact on academic achievement and that its effectiveness is affected by various moderators. For example, Topping (2005) emphasized that 12 organizational dimensions should be considered when planning peer learning (i.e., peer tutoring and cooperative learning), including participant characteristics, training, and tutoring parameters. Similarly, Topping (1996) suggested that 10 organizational dimensions should be considered when planning peer tutoring in higher education, such as tutor and tutee characteristics, pairing methods, and reciprocal peer tutoring. McMaster et al. (2006) also postulated that the success of peer-assisted learning for reading, which is a type of classwide peer tutoring, depends on the inclusion of key features such as training, fidelity checks, and tutoring duration.

Although these traditional literature reviews have attempted to synthesize studies to draw consistent conclusions, such reviews rely on qualitative summaries or "vote counting" of results based on their statistical significance (Lipsey & Wilson, 2001). Such reviews are strongly affected by the subjective opinions, the sample sizes, and selection procedures.

Meta-analysis allows for the systematic reviewing of empirical research to identify relationships among study characteristics and outcomes (Durlak, 1995). This methodology synthesizes the findings of different studies to observe effect sizes across these studies (Rosenthal, 1984). Meta-analysis provides several advantages over a traditional literature review. For example, rather than classifying results into positive or null categories based on their statistical significance or conducting qualitative summaries, as occurs in traditional literature reviews, meta-analysis permits the evaluation of the direction and magnitude of the effects observed in each study, the distribution of effects across studies, and the computation of the average effect for a group of studies (Durlak & Lipsey, 1991; Lipsey & Wilson, 2001). Meta-analysis explicitly and

This article was published Online First October 20, 2014.

Correspondence concerning this article should be addressed to Kim Chau Leung, Department of Special Needs and Counselling, Hong Kong Institute of Education, 10 Lo Ping Road, Taipo, NT, Hong Kong. E-mail: ckcleung@ied.edu.hk

clearly states the sampling method, selection criteria, procedures for identifying and retrieving eligible studies, characteristics, and effect sizes of the studies (Lipsey & Wilson, 2001). Hence, meta-analyses are more readily replicated and criticized based on their assumptions, judgments, and procedures (Durlak & Lipsey, 1991). Unlike traditional reviews, this process reduces reviewer bias throughout the investigation and allows conclusions to be drawn based on strong scientific evidence rather than expert opinion (Durlak, 1995). Hence, the results obtained through meta-analysis are more convincing than are those obtained by traditional reviews.

Meta-analysis can be coded to address numerous study characteristics and examine a wide range of relationships and interactions among these variables using multivariate techniques (Durlak & Lipsey, 1991). Furthermore, it permits policymakers to examine program effectiveness and identify the moderating variables that affect program effectiveness. Hence, meta-analysis allows policymakers to evaluate which features of an intervention constitute the best practices for optimal effectiveness of the intervention. In the next section, several meta-analyses of the effects of peer tutoring on academic achievement are discussed to reveal their limitations and explain the compelling need for the present investigation.

Previous Meta-Analyses of Peer Tutoring: Contributions and Limitations

Cohen, Kulik, and Kulik's (1982) meta-analysis set the stage for the present investigation. This classic study evaluated the effects of peer tutoring on the academic achievement of tutees at the elementary and secondary school levels. It provides evidence that peer tutoring positively affects tutee achievement. The subject area, student characteristics, intervention features (including grade and ability levels of the students), and duration of tutoring were found to moderate the effects of tutoring on achievement. Although this meta-analysis was comprehensive, it is now dated, because the studies that it examined were conducted prior to 1980. In addition, not all of the tutoring provided was peer tutoring. This meta-analysis included all types of tutoring, including teacher-led and other adult-led tutoring. Hence, the results are confounded by the inclusion of different types of tutoring. Although the study evaluated the effects of tutoring considering certain student characteristics and intervention features, more rigorous statistical procedures could be employed in current meta-analytical research, such as homogeneity analyses. Notably, this study was conducted prior to the development of many methods in meta-analysis and, therefore, did not correct for the effects of small samples, weigh the effect sizes by sample sizes, or consider sample size outliers.

Subsequent meta-analyses (e.g., Cook, Scruggs, Mastropieri, & Casto, 1985; Mathes & Fuchs, 1994; Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003) also have indicated that peer tutoring has a positive impact on academic achievement. Moreover, meta-analyses have shown that some program features, such as subject content, student characteristics, and intervention features, moderate the effects of peer tutoring. However, some of these studies are also dated. For example, the meta-analytic reviews conducted by Cook et al. (1985) and Mathes and Fuchs (1994) did not capture methodological advances, such as correcting for sample size when calculating the magnitude of the effect, utilizing weighting procedures that account for sample size, selecting the appropriate unit of analysis, winsorizing outliers, and conducting homogeneity analyses to examine moderator variables and group differences. Moreover, recent meta-analyses of peer tutoring in which some of these advances were adopted were still plagued by methodological flaws. For example, the meta-analysis conducted by Rohrbeck et al. (2003) was restricted to certain populations (such as elementary-school children), confounded peer tutoring with other modes of learning (such as cooperative learning), failed to estimate both fixed and mixed effects models for effect size, and did not employ current methods for addressing publication bias.

Hence, there is a compelling need to conduct an updated, comprehensive meta-analysis to establish the effects of peer tutoring, unconfounded with other types of peer-assisted learning, on academic achievement that reflects advances in meta-analysis methodology and a wide range of participants and subjects. Thus, a primary aim of the present investigation was to address these issues.

In general, the parameters examined in the present investigation were adopted from the meta-analyses mentioned earlier. However, some program parameters that have not yet been examined were investigated, including the educational levels of both the tutee and tutor, frequency of tutor training, length of each session, type of reward, mode of assigning a dyad pair, and parental involvement.

Some theoretically based program parameters that have not been validated in previous studies were also examined in the present investigation, including the type of peer tutoring and dyad gender as explained by role theory and the formation of teams competing for reward and new competing team formation as explained by interdependent group contingencies.

Theoretically Based Program Parameters

According to role theory, people behave according to the roles assigned to them. Tutors perceive themselves as academically competent and display positive attitudes toward school. In contrast, acting as the role of the tutee would cause tutees to perceive themselves as less competent and inferior (Bierman & Furman, 1981). Hence, the tutee may dislike or envy the tutor due to their unequal social status. Falchikov (2001) suggested that adopting reciprocal peer tutoring could reduce social inequality and tutee hostility toward tutors because each student takes a turn as tutor and tutee. Hence, it was hypothesized that both cross-age peer tutoring and same-age reciprocal peer tutoring would produce greater effects than would same-age nonreciprocal peer tutoring.

Robinson et al. (2005) applied role theory to explain why mixed-sex pairs are not as effective as same-sex pairs. First, students are concerned with how they perform gender role-related behaviors. For example, female tutors in mixed-sex pairs display negative responses to tutoring (Fogarty & Wang, 1982). Female tutors may find it uncomfortable to perform the role of the tutor because it confers greater authority and superiority to the female than to the male tutee (Eagly, Wood, & Diekman, 2000), which is inconsistent with traditional gender roles. Additionally, male tutees may be uneasy with performing the subordinate role of tutee when a female tutor occupies the superior status. Gender roles also negatively affect social processes in mixed-sex pairs. Topping and Whiteley (1993) revealed that mixed-sex pairs performed less effectively than same-sex pairs did, regardless of whether the tutor was male or female. This result occurs because students are preoccupied with their gender roles and distracted from learning.

Underwood, Underwood, and Wood (2000) revealed that mixed-sex pairs of 9- to 11-year-old students produced less discussion of the learning material, fewer on-task behaviors, and less enjoyment during computer-based problem-solving activities compared with same-sex pairs.

Second, mixed-sex pairs may produce stereotype threats, which negatively affect tutee performance. For example, girls are not expected to perform as well as boys do in mathematics (Eccles, Jacobs, & Harold, 1990; Swim, 1994). Females are aware of this gender stereotype and confirm it. Numerous studies reveal that the mathematics performance of females from the elementary to college levels are negatively affected by this gender stereotype threat (e.g., Ambady, Shih, Kim, & Pittinsky, 2001; Marx & Roman, 2002). Girls who are tutored by boys may feel very uneasy with disproving this stereotype in mathematics (Robinson et al., 2005); hence, their performance is hindered.

It was hypothesized that same-sex dyads would produce greater effects than would mixed-sex dyads.

Regarding the concept of interdependent group reward contingencies, people are rewarded based on the joint effort of all team members, as commonly occurs in cooperative learning. Slavin (1990) found that interdependent group reward contingencies are more effective than independent reward contingencies because motivation is greatly increased when people are mutually dependent on each other and pursue mutual goals to complete a task (Johnson, Maruyama, Johnson, Nelson, & Skon, 1981; Wentzel, 1999). People support each other and emphasize joint effort when they work toward a common goal (Johnson et al., 1981; Slavin, 1990). Because some studies of peer tutoring (e.g., Dion et al., 2011; Sáenz, Fuchs, & Fuchs, 2005) have incorporated interdependent group reward contingencies, it is interesting to evaluate whether team formation to compete for rewards in peer tutoring is more effective than peer tutoring without teams competing for rewards. In other words, this analysis helps evaluate the effectiveness of incorporating elements drawn from other peer-assisted interventions, such as cooperative learning, into peer tutoring.

In sum, in the present investigation, the following hypotheses were tested:

> *Hypothesis 1:* Both cross-age peer tutoring and same-age reciprocal peer tutoring produce greater effects than does same-age nonreciprocal peer tutoring.
>
> *Hypothesis 2:* Same-sex dyads produce greater effects than do mixed-sex dyads.

In addition, the present investigation explored whether forming teams to compete for rewards produces greater effects than peer tutoring without competing teams does, consistent with interdependent group reward contingencies. It would also be interesting to evaluate the effect of the regular formation of new teams on the effectiveness of interdependent group reward contingencies.

Aims of the Present Investigation

The aim of the present investigation was to synthesize previous research pertaining to peer tutoring and address the limitations of previous meta-analytic research by including studies that examined wide ranges of subject content and participants and adopting current methodological approaches to meta-analysis to meet the following aims:

1. Critically evaluate the overall effectiveness of peer tutoring on academic achievement;
2. Identify and compare the moderators of the effectiveness of peer tutoring on academic achievement;
3. Validate the predicted effects of certain program parameters based on role theory and interdependent group contingencies, as discussed in the preceding sections; and
4. Propose a preliminary empirical model of the crucial determinants of best practices for peer tutoring on achievement.

Method

Selection Criteria, Procedures, and Sample of Studies

Search for the literature covered various sources. First, key terms, including "peer tutoring," "peer tutee," "peer tutor," "tutoring," "tutor," and "tutee," were used to search two main online databases: PsycINFO and Educational Resources Information Center (ERIC). Second, relevant studies from previous meta-analytic studies and reference lists from the review articles of peer tutoring were examined. Finally, a manual search was conducted on journals that publish peer-tutoring articles including *Journal of Educational Psychology, American Educational Research Journal, Elementary School Journal, Journal of Applied Behavior Analysis, Journal of Experimental Education, School Psychology International, School Psychology Review, School Psychology Quarterly, Higher Education, Contemporary Educational Psychology, Educational Psychology, Educational Studies, Educational Research, Journal of Special Education, Remedial and Special Education, Education & Treatment of Children, Journal of Learning Disabilities, Behavioral Disorders, Exceptional Children, Learning Disabilities Research and Practice,* and *Reading Research Quarterly.* The following criteria were set for the eligibility of studies: (a) the study was peer-reviewed and published in 2012 or before; (b) peer tutoring occurred in a school setting; (c) participants were kindergarten, elementary, secondary, college, or university students; (d) targeted subject matter was academic achievement; (e) the outcome data available in the article were amenable to the computation of effect sizes; and (f) the experimental design included treatment-comparison (quasi-experimental) or treatment-control (true experimental) group design. Conference presentations, dissertations, reports, and book chapters were excluded because of occurrence of bias. As Ferguson and Brannick (2012) revealed, authors of meta-analyses themselves were overrepresented in unpublished studies acquired compared with published studies based on examination of a sample of 91 recent meta-analyses published in American Psychological Association and Association for Psychological Science journals and the methods adopted in these studies to identify and control for publication bias. It indicated that including unpublished studies in the meta-analyses increases rather than decreases bias, probably due to selection bias in unpublished literature searches. Moreover, non-English materials were excluded because of the limited language abilities of the author. In this study, 72 articles reporting sufficient data for the computation of effect size were retained for further analysis. A complete list of articles can be obtained from the author.

Coding Procedures

The coded features had been utilized in a previous meta-analysis of tutoring (Cohen et al., 1982) and provided a basis for developing a coding sheet (see the Appendix in the online supplemental material). Specifically, this coding schema included (a) report information, (b) participants' characteristics, (c) methodology, (d) intervention features, and (e) outcomes assessment.

Two coders used the coding sheet to code the articles. To ensure a common understanding of the items on the coding sheet, two pilot coding sessions were held to discuss any disparities on the coding sheet. After consensus had been reached, each coder performed his or her coding separately. While the author coded all 72 studies, the second coder coded 36 randomly selected studies. Initial interrater reliability was calculated using percentage of agreement and kappa coefficients, where appropriate, for these studies. The average percentage agreement and kappa coefficient for the variables were 97.0% and 0.96%, respectively. Any disparity in coding was resolved with thorough discussion until consensus was reached.

Unit of Analysis and Shifting Approach

Because independent samples were the primary unit of analysis, each study contributed one independent sample to the analysis; hence, one effect size was calculated. However, when there were subgroups within a study, a shifting unit of analysis approach was utilized to determine which constituted an independent unit of analysis (Cooper, 1998). First, for each subgroup within a single study, each effect size was coded as if it were an independent estimate of the outcome. For example, if a study reported findings on achievement separately for different subgroups, such as low achiever, average achiever, and high achiever, these effect sizes were calculated separately. Then, when counting the overall impact of peer tutoring on achievement, only one effect size was calculated for that study by averaging the effect sizes across these three subgroups. However, when evaluating the moderating effect of peer tutoring on achievement based on different ability levels of participants, three effect sizes were calculated.

Computation of Standardized Effect Size

The standardized effect size was calculated by dividing the difference between the treatment and control/comparison group means by the pooled standard deviation of the two groups (Hedges, 1981). Since effect size is positively biased with small samples (Hedges & Olkin, 1985), standard procedure for correcting such bias was adopted.

Data Analyses

Mean effect size and variance. The Comprehensive Meta-Analysis software program (Version 2.0; Borenstein, Hedges, Higgins, Rothstein, 2005) and SPSS Macro (see Lipsey & Wilson, 2001) were used to compute the mean effect size, variance and 95% confidence interval estimates for both the fixed and mixed effects models. Method of moments was used for estimation of variance in mixed effects model (Hedges & Vevea, 1998). A nonzero 95% confidence interval indicates that the mean effect was significant (Hunter & Schmidt, 1990) and that a smaller confidence interval can be found in fixed-effects model (Hedges & Vevea, 1998).

Homogeneity tests, fixed and mixed effects models, and I^2 index. Homogeneity tests were adopted to evaluate whether the mean values of various effect sizes all estimated the same population effect size (Hedges, 1982; Rosenthal & Rubin, 1982). A homogeneous distribution of effect sizes suggests that the dispersion of effect sizes around their mean is accounted for by subject-level sampling error alone. Conversely, in heterogeneous condition, each effect size does not estimate the same population effect size, and the variability of the effect sizes would be explained by not only subject-level sampling error but also study-level sampling error associated with different study characteristics, such as methodology (Lipsey & Wilson, 2001).

Fixed and mixed effects models are two important methods for analyzing the effect size in the present investigation. A fixed effects model assumes that any variance in effect size is due to subject-level sampling error (Hedges & Vevea, 1998), whereas a mixed effects model suggests that study-level variance are further sources of errors (Lipsey & Wilson, 2001; Raudenbush, 1994). Hence, a homogeneous condition indicates that assumptions based on the fixed-effects model are supported and that only the subject-level sampling error is included. A heterogeneous condition suggests that the assumptions based on the fixed effects model are rejected, that both the subject-level sampling error and the study-level sampling error associated with different study characteristics should be considered, and, hence, that the mixed effects model should also be examined.

Homogeneity was evaluated utilizing a homogeneity test, which is based on a Q statistic, distributed chi-square (Hedges & Olkin, 1985). A significant Q statistic rejects the null hypothesis of homogeneity and suggests a heterogeneous condition.

However, the Q statistic does not provide information on the degree of heterogeneity. Higgins and Thompson (2002; Higgins, Thompson, Deeks, & Altman, 2003) suggested adopting the I^2 index to assess the extent of heterogeneity by dividing the difference between the value of Q statistic and its degree of freedom ($k - 1$) by the Q value and multiplying by 100 (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). The I^2 index indicates the percentage of total variance in a set of effect sizes resulted from heterogeneity. An I^2 index of 25, 50, or 75 tentatively suggests low, medium, or high heterogeneity, respectively (Higgins et al., 2003; Higgins & Thompson, 2002).

Test for moderators of effects. Homogeneity tests were conducted to evaluate whether each set of effect sizes estimated the same population effect size (Hedges, 1982; Rosenthal & Rubin, 1982). They help evaluate whether the variability in the effect size within a particular set is accounted for by the sampling error alone or by a special coded feature of the effect size of the set. Homogeneity was examined using a between-group homogeneity statistic, Q_B, and a within-group homogeneity statistic, Q_W (Hedges & Olkin, 1985). The statistic Q_B is analogous to the main effect in an analysis of variance design, and it examines the differences between groups of effect sizes. A significant Q_B indicates that the grouping variable is a significant moderator of outcome and that the average effect size differs between subgroups. However, Q_w is used to test the within-group effect, and a nonsignificant Q_w suggests that the studies can be grouped into homogeneous subgroups. There-

fore, the effect sizes under this subgroup were consistent across the studies (Lipsey & Wilson, 2001).

As discussed in the preceding section, in a fixed effects model analysis, variance in effect size is due to participant-level sampling error (Hedges & Vevea, 1998). In a mixed effects model analysis, study-level variance is an additional source of error (Lipsey & Wilson, 2001; Raudenbush, 1994). Hence, a fixed-effects analysis permits drawing inferences from studies that are observed in the present investigation, whereas a mixed-effects analysis allows inferences to be drawn from studies beyond the present sample (Hedges & Vevea, 1998; Valentine, DuBois, & Cooper, 2004) and thus permits any inferences that are drawn to be generalized to a larger population. In the present investigation, it is also interesting to draw inferences about all peer tutoring interventions beyond the studies included in this study because there is considerable variation in the study characteristics, such as procedures and intervention parameters, which may contribute to study-level errors. Hence, it is appropriate to adopt mixed effect models in addition to fixed effects models.

Additionally, when homogeneity is rejected, a fixed effects model may underestimate error variance, while a mixed effects analysis may overestimate it (Overton, 1998). Given both this tendency and different inferences about the intervention effects produced by these two approaches, the present investigation adopts both fixed and mixed-effects models, which previous intervention studies have not considered. This is also in line with the current trend of adopting multiple approaches (e.g., O'Mara, Marsh, Craven, & Debus, 2006; Patall, Cooper, & Robinson, 2008; Valentine et al., 2004). Moreover, in addition to Q statistics, the I^2 index was adopted to assess the extent of heterogeneity, as mentioned previously.

Publication Bias

Because only published articles were included in this meta-analysis, the effect size would be overestimated because published materials are more likely to present significant findings and findings that support the hypothesis (Lipsey & Wilson, 2001). Various methods have been developed to evaluate whether this publication bias occurred. Traditionally, a Fail-Safe N analysis (Rosenthal, 1979) was conducted to estimate the number of unpublished studies with no significant effect size that would be included to reject the overall effect size in this investigation to a trivial level (Orwin, 1983).

However, Fail-Safe N analysis does not estimate the number of missing studies (Soeken & Sripusanapan, 2003). McDaniel, Rothstein, and Whetzel (2006) recommended the use of the defensible trim-and-fill method (Duval & Tweedie, 2000a, 2000b) over the traditional Fail-Safe N analysis because it can provide a reasonable calculation of the missing studies (Duval & Tweedie, 2000a, 2000b; Pham, Platt, McAuley, Klassen, & Moher, 2001). Hence, only the trim-and-fill approach was reported in the present investigation.

The trim-and-fill method is a nonparametric iterative method that attempts to adjust the effect of missing studies due to publication bias on the observed distribution of effect sizes in the funnel plot (Duval & Tweedie, 2000a, 2000b). It involves removing the most extreme small studies from the positive side of the funnel plot (asymmetric part of the funnel) and re-estimating the effect size in each iteration until the symmetry of the funnel plot is restored. As a result, the effect size and variance of the effects are smaller, and a narrow confidence interval is found (Borenstein, 2005; Duval & Tweedie, 2000a, 2000b).

As recommended by Duval (2005) and Sutton (2005), the trim-and-fill method was applied to both the fixed and mixed effects models in the present investigation to compare the results of the two approaches.

Results

Descriptive Features of Peer Tutoring

The process of identifying studies is depicted in the flow diagram displayed in Figure 1. An initial search revealed 15,821 articles. After excluding studies (n = 15,517) that were likely irrelevant based on the title, abstract, and sources (e.g., book chapters and conference presentations), 304 remained. After excluding studies that did not meet the inclusion criteria, 72 articles published in English were included in this meta-analysis. The characteristics of the 72 studies are listed in Tables 1–5. Most studies were conducted during the periods 1990–1999 (41.7%; n = 30) and 2000–2012 (37.5%; n = 27).

The mean age of the tutees was 12.66 years (SD = 6.20), while that of the tutors was 13.06 years (SD = 6.55). Because most of the studies (75.0%, n = 54) adopted same-age peer tutoring, the mean ages of the tutors and tutees were similar, and there was a very high correlation (.99, $p < .001$).

The mean number of tutees was 68.01 (SD = 88.10), and the numbers ranged from three to 494 tutees. Conversely, the mean

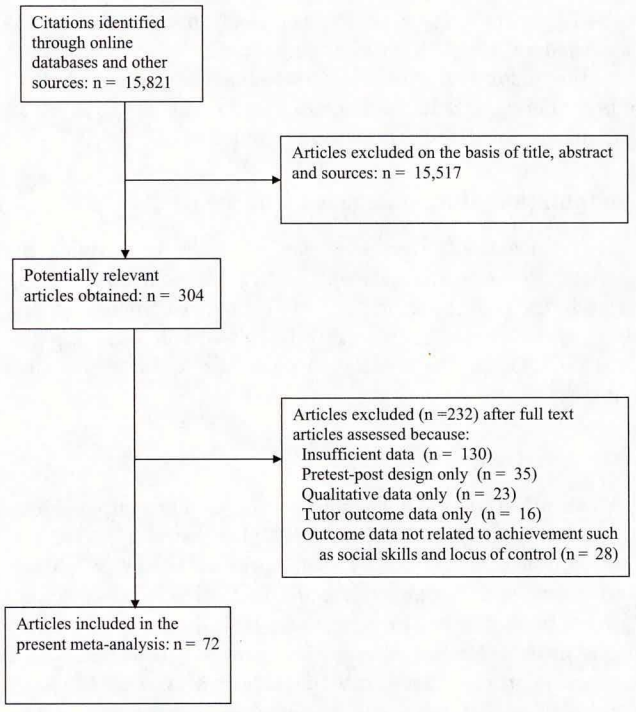

Figure 1. Flowchart of systematic literature search.

Table 1
Descriptive Statistics of Participant Parameters at the Study Level (N = 72)

Variable	*n*	%
Education level of tutee		
Kindergarten	2	2.8
Elementary school	48	66.7
Secondary school	12	16.7
College or university	9	12.5
Both elementary and secondary	1	1.3
Education level of tutor		
Kindergarten	2	2.8
Elementary school	46	63.9
Secondary school	13	18.1
College or university	10	13.9
Both elementary and secondary schools	1	1.3
Academic ability of tutee		
Low	28	38.9
Average	0	0
High	4	5.6
Unspecified group	17	23.6
Mixed[a,b]	23	31.9
Academic ability of tutor		
Low	26	36.1
Average	0	0
High	6	8.3
Unspecified group	18	25.0
Mixed[a,b]	22	30.6
Socioeconomic status (SES) of participants		
Low	12	16.7
Middle	2	2.8
High	0	0
Mixed	12	16.7
Unspecified	45	62.5
Two distinct SES groups: both low and high SES[b]	1	1.3
Minority of participants		
≤50%	22	30.6
>50%	17	23.6
Unspecified group	33	45.8

[a] Seven of these studies had three distinct ability groups: low, average and high, and six had two distinct ability groups: low and average. [b] Separate effect size was computed for each subgroup.

number of tutors was 52.91 (SD = 52.08), with numbers ranging from three to 277 tutors. Of the studies, 45.8% (n = 33) reported the gender of the tutees. There were 52.3% male and 47.7% female tutees in total. Similarly, of the studies, 43.1% (n = 31) reported the gender of tutors. There were 53.7% male and 46.3% female tutors overall.

Regarding the instruments for assessing the outcomes, they were categorized into standardized or unstandardized tests. Standardized tests involve using a standard set of tutoring materials, a standard administration procedure, and a standard scoring procedure (Elbaum, Vaughn, Hughes, & Moody, 2000), such as the Comprehensive Reading Assessment Battery (Fuchs, Fuchs, & Hamlett, 1989); conversely, unstandardized tests, such as those developed by the project staff or teacher, do not have these features. In the present investigation, most of the measures for evaluating the achievement of mathematics and reading were standardized tests, whereas most are unstandardized tests for subjects (e.g., physical education, arts) other than mathematics and reading.

Table 2
Descriptive Statistics of Methodology Parameters at the Study Level (N = 72)

Variable	*n*	%
Structured tutoring		
Yes	49	68.1
No	19	26.4
Mixed type[a]	4	5.5
Parent involvement		
Yes	2	2.8
No	69	95.8
Mixed type[a]	1	1.4
Same-gender dyad		
Same gender	14	19.4
Mixed gender	58	80.6
Control for author bias[b]		
Yes	39	54.2
No	30	41.7
Mixed type[c]	3	4.1
Reward type		
Earning points	12	16.7
Points converted into tangible items	3	4.1
Tangible items	6	8.4
No reward specified	50	69.4
Mixed type (tangible reward and no reward)	1	1.4
Formation of competing team for group reward		
Yes	12	16.7
No	60	83.3
Fidelity check		
Yes	39	54.2
No	33	45.8
Assignment of pairing		
Voluntary basis	2	2.8
Was assigned	70	97.2
Regular formation of new competing team for team reward		
Yes	7	58.3
No	5	41.7
Frequency of tutor training[d] (number of sessions)		
≤2.5 sessions per week	17	23.6
>2.5 sessions per week	18	25.0
Unspecified group[d]	37	51.4
Tutor training		
Yes	49	68.1
No	17	23.6
Mixed mode[a]	6	8.3
Length of tutor training[e] (minutes per session)		
≤45 min per session	18	25.0
>45 min per session	15	20.8
Unspecified group[e]	39	54.2
Random assignment		
Yes (i.e. experimental design with control group)	52	72.2
No (i.e. quasi-experimental design with comparison group)	19	26.4
Mixed mode[a]	1	1.4
Type of peer tutoring		
Same-age reciprocal peer tutoring	39	54.2
Same-age nonreciprocal peer tutoring	12	16.7
Cross-age peer tutoring	16	22.2
Both same-age nonreciprocal and cross-age peer tutoring	2	2.8
Both same-age reciprocal and nonreciprocal peer tutoring[a]	3	4.1

[a] Separate effect size was computed for each subgroup. [b] Standardized test used. [c] Two of these studies had two distinct groups; separate effect size computed for each. [d] Median = 2.5 sessions per week. [e] Median = 45 min per session.

Table 3
Descriptive Statistics of Intervention Parameters at the Study Level (N = 72)

Variable	*n*	%
Duration of tutoring (no. of weeks)		
≤10 weeks	32	44.5
>10 weeks	28	38.9
Unspecified group (median = 10 weeks)	12	16.6
Frequency of each tutoring session (no. of session per week)		
≤2.5 sessions per week	29	40.3
>2.5 sessions per week	25	34.7
Unspecified group (median = 2.5 sessions per week)	18	25.0
Length of each tutoring session (minutes per session)		
≤30 min per session	29	40.3
>30 min per session	25	34.7
Unspecified group (median = 30 min per session)	18	25.0
Total dosage of tutoring (no. of hour)		
≤16.25 hours	22	30.5
>16.25 hours	21	29.2
Unspecified group (median = 16.25 hr)	29	40.3

Profile of Effect Size

There were 72 studies regarding the achievement of tutees. Some of these studies reported findings separately for different subgroups (e.g., low, mixed, and high ability). Hence, each of these studies contributed more than one effect size. However, when independent samples were the primary units of analysis used to calculate the overall effect size across 72 studies, each study contributed one independent sample to the analysis, and only one effect size was calculated by averaging the effect sizes of the subgroups within each study. Hence, analyses of the effect sizes entailed 367 effect sizes for achievement, whereas for independent samples, only 72 effect sizes for achievement were used. Figure 2 shows the distribution of effect sizes.

It was important to examine the distribution of the unweighted effect sizes to identify any outliers because extreme values (outliers) exert undue influence on the values of effect sizes such as the means and variances of effect sizes (Durlak & Lipsey, 1991). Positive outliers were defined as effect sizes that were more than three interquartile ranges (IQ) beyond the third quartile, Q3 (75th percentile), whereas negative outliers were referred to as effect sizes that were more than three IQs below the first quartile, Q1 (25th percentile) in the box plot, based on Tukey's definition (Tukey, 1977). IQ is the range between Q3 and Q1; hence, Q1 and Q3 should be identified first for calculating the IQ. After the outliers were revealed, they were winsorized by setting their values to three IQs beyond the 75th percentile for positive outliers and three IQs below the 25th percentile for negative outliers. This procedure reduced the undue impact of these outliers on the subsequent effect size analyses, but their large sizes were still accounted for (Durlak & Lipsey, 1991; Tabachnick & Fidell, 2001). In the present investigation, there were only two positive outliers with extreme effect sizes of 3.68 (Limbrick, McNaughton, & Glynn, 1985) and 5.47 (Mackiewicz, Wood, Cooke, & Mazzotti, 2010). After these two extreme effect sizes were winsorized, the unweighted mean effect size for the 72 independent samples as the unit of analysis was 0.59 (SD = 0.58).

Examining the distribution of the sample size to check for outliers was also necessary because the weighting of effect size was based on sample size, and hence, extremely large sample sizes would have had undue effects on the findings. Following the previously described procedures, only two studies of achievement with extreme sample sizes of 530 (Topping et al., 2011) and 705 (Topping, Thurston, McGavock, & Conlin, 2012) met the criteria as outliers as defined by Tukey (1977). After winsorizing these two positive outliers, the average sample size of the studies included in this investigation was 91.83 (SD = 110.75).

The weighted mean effect size for the 72 independent samples included in the present investigation of achievement was significant in both the fixed effects model (d = 0.39, $p < .001$; 95% confidence interval [CI] [0.36, 0.41]) and mixed effects model (d = 0.47, $p < .001$; 95% CI [0.39, 0.66]).

The homogeneity statistics (Q [71, k = 72] = 589.46, $p < .001$, I^2 = 88%) for the fixed effects model were significant, indicating that homogeneity assumptions regarding the effect size distribution in the fixed effects model were rejected and high heterogeneity (88%) occurs and that cautions should be taken when generalizing the results of the fixed effects model beyond the studies examined in the present investigation. In this connection, analyses based on mixed effects model were necessary in addition to the fixed effects model. Moreover, the heterogeneity analyses suggested that it was important to examine the moderators of effect sizes that would affect the effect size distribution.

Regarding publication bias, the trim-and-fill models was conducted in the present investigation. It was revealed that the unweighted mean effect size was 0.59 (SD = 0.58) and the weighted mean effect size was 0.39 and 0.47 for the fixed and mixed effects models, respectively, before trimming. As shown in the funnel plot, it was unsymmetrical due to certain effect sizes in the positive side of the funnel (see Figure 3). Hence, it indicated that a certain degree of publication bias occurred. After trimming, the weighted mean effect size was d = 0.26 (95% CI [0.24, 0.28]) for the fixed effects model and d = 0.37 (95% CI [0.29, 0.45]) for the mixed effects model. Because the confidence interval did not include zero, the effect size was significantly different from zero at the 5% level of significance. Hence, the effect of peer tutoring on tutees' academic achievements is significant. The positive effect indicated that there were greater posttest scores in the treatment groups than in the comparison/control groups.

Table 4
Descriptive Statistics of Outcome Parameter at the Study Level (N = 72)

Target subject matter on achievement	*n*	%
Mathematics	17	23.6
Reading	27	37.5
Language	2	2.8
Science and technology	6	8.3
Physical education	4	5.6
Arts	4	5.6
Psychology	4	5.6
Education	1	1.4
Miscellaneous subject	3	4.1
Both reading and language[a]	1	1.4
Both mathematics and reading, language[a]	3	4.1

[a] Separate effect size was computed for each subgroup.

Table 5
Descriptive Statistics of Instrument Adopted for Assessing Outcome Parameter at the Study Level (N = 72)

Subject/instrument	Good reliability	Good criterion validity	No. of studies
Reading			
Standardized test			
1. Comprehensive Reading Assessment Battery (CRAB; Fuchs, Fuchs, & Hamlett, 1989)	Yes	Yes	8
2. Woodcock Reading Mastery Test (WRMT; Woodcock, 1973)	Yes	Yes	3
3. Test of Early Reading Ability (TERA; Reid, Hresko, & Hammill, 2001)	Yes	Yes	2
4. Stanford Diagnostic Reading Test (SDRT; Karlsen & Gardner, 1986)	Yes	Yes	1
5. Stanford Achievement Test (SAT; (Gardner, Rudman, Karlsen, & Merwin, 1982)	Yes	Yes	1
6. Woodcock–Johnson Achievement Test (Woodcock, McGrew, & Mather, 2001)	Yes	Yes	1
7. Neale Analysis of Reading Ability (Neale, 1997)	Yes	Yes	3
8. Burt Word Reading Test (Vernon, 1969)	Yes	Yes	1
9. Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashottee, 1999)	Yes	Yes	1
10. Comprehensive Test of Phonological Awareness (CTOPP; Wagner, Torgesen, & Rashottee, 1999)	Yes	No	1
11. Index of Reading Awareness (IRA; Jacobs & Paris, 1987)	Yes	Yes	1
12. Reading Strategy Use Scale (RSU; Pereira-Laird & Deane, 1997)	Yes	Yes	1
13. Metropolitan Achievement Test (Prescott, Balow, Hogan, & Farr, 1978)	Yes	Yes	1
14. Suffolk Reading Test (Hagley, 2002)	Yes	Yes	1
15. Wide Range of Achievement Test (Jastak & Jastak, 1965)	Yes	Yes	2
Researcher-developed with good reliability and content validity reported	Yes	NA	1
Researcher-developed with relevant content validity reported but without reliability reported	Not reported	NA	4
Subject test provided by teacher with relevant content validity but without reliability reported	Not reported	NA	1
Mathematics			
Standardized test			
1. Comprehensive Mathematics Test (Stecker, Fuchs, & Hamlett, 1992)	Yes	Yes	1
2. Math Operations Test–Revised (Fuchs, Fuchs, Hamlett, & Stecker, 1991)	Yes	Yes	2
3. Metropolitan Achievement Test (Prescott, Balow, Hogan, & Farr, 1978)	Yes	No	1
4. Stanford Diagnostic Mathematics Test (SDMT; Beatty, Madden, Gardner, & Karlsen, 1986)	Yes	Yes	2
5. Curriculum-Based Computation Test (Thurber, Shinn, & Smolkowski, 2002)	Yes	No	5
6. Standard Arithmetic Test (Bar-Eli & Raviv, 1982)	Yes	Yes	1
7. Woodcock–Johnson Mathematics Achievement Test (Woodcock et al., 2001)	Yes	Yes	1
8. Key Math Diagnostic Arithmetic Test (Roach, Paolucci-Whitcomb, Meyers, & Duncan, 1983)	Yes	Yes	1
9. California Achievement Test (Sprinthall & Scott, 1989)	Yes	Yes	1
10. Stanford Achievement Test (SAT; Gardner, Rudman, Karlsen, & Merwin, 1987)	Yes	Yes	1
11. Wide Range Achievement Test (Jastak & Jastak, 1965)	Yes	Yes	2
Researcher-developed with good reliability and content validity reported	Yes	NA	1
Researcher-developed with relevant content validity reported but without reliability reported	Not reported	NA	2
Subject test provided by teacher with relevant content validity but without reliability reported	Not reported	NA	2
Other subjects			
Standardized test			
1. Metropolitan Achievement Test (Prescott, Balow, Hogan, & Farr, 1978)	Yes	Yes	2
2. Wide Range Achievement Test (Jastak & Jastak, 1965)	Yes	Yes	2
3. Writing Expression of Woodcock Johnson–Revised (Woodcock et al., 2001)	Yes	Yes	1
4. Test of Written Language (TOWL; Hammill & Larsen, 1988)	Yes	Yes	1
Researcher-developed with good reliability and content validity reported	Yes	NA	9
Researcher-developed with relevant content validity reported but without reliability reported	Not reported	NA	1
Subject test provided by teacher with relevant content validity but without reliability reported	Not reported	NA	11

Test for Moderators of Effects for the Fixed Effects Model

Participant parameters. In essence, the educational levels of the participants significantly moderated achievement outcomes, yielding high heterogeneity ($Q_B = 34.79$, $p < .001$, $I^2 = 91\%$). Studies of tutees from secondary schools displayed the largest effect sizes ($d = 0.52$), followed by college or university ($d = 0.43$), elementary school ($d = 0.34$), and kindergarten ($d = 0.25$) students (see Table 6). Since number of studies in kindergarten group was too small ($n = 2$), it was important to examine whether removal of this group would have significant impact on the result. However, removal of kindergarten group did not affect the result ($Q_B = 29.65$, $p < .001$, $I^2 = 93\%$), and same descending order of effect sizes was revealed as shown above.

The academic ability of the tutee was a significant moderator of achievement outcome, which yielded a high degree of heterogeneity ($Q_B = 27.35$, $p < .001$, $I^2 = 89\%$). Studies of tutees with high ability levels displayed the largest effect sizes ($d = 0.44$), followed by low ($d = 0.41$), average ($d = 0.29$), and mixed ($d = 0.26$) ability levels.

The academic ability of the tutor was a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 27.08$, $p < .001$, $I^2 = 89\%$). Studies of tutors with low ability levels displayed the largest effect sizes ($d = 0.43$), followed by high ($d = 0.39$), average ($d = 0.29$), and mixed ($d = 0.27$) ability levels. As mentioned previously, similar patterns of the distributions of effect sizes were found for the academic abilities of tutees and tutors because most of the studies adopted same-age peer tutoring.

```
Frequency   Stem &  Leaf

   1.00 Extremes    (-.78)
   2.00      -3 .  00
    .00      -2 .
   4.00      -1 .  0355
   9.00      -0 .  123335789
  11.00       0 .  00011134899
  12.00       1 .  112445677889
  13.00       2 .  0013566677889
  18.00       3 .  000122223445566678
  18.00       4 .  001123334455555799
  11.00       5 .  01233456678
   8.00       6 .  24557889
  11.00       7 .  00013456699
   8.00       8 .  11134449
   3.00       9 .  668
   7.00      10 .  1223569
   6.00      11 .  044578
    .00      12 .
   3.00      13 .  478
   1.00      14 .  3
   2.00      15 .  59
   4.00 Extremes    (2.42), (2.77), (2.79), (3.52)
```

Figure 2. Stem and leaf plot showing distribution of effect sizes.

Minority of participants was a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 3.89$, $p < .05$, $I^2 = 74\%$). Studies with 50% or fewer participants from minority groups displayed larger effect sizes ($d = 0.40$) than did those with more than 50% minority participants ($d = 0.33$).

The socioeconomic status (SES) of participants did not significantly moderate effect size ($Q_B = 5.73$, *ns*, $I^2 = 65\%$). More specifically, similar effect sizes were found regardless of SES. Since the number of studies in middle SES was too small ($n = 2$), it was important to examine whether removal of this group would have significant impact on the result. However, removal of the middle SES group did not affect the result ($Q_B = 3.33$, *ns*, $I^2 = 70\%$), and there was no significant difference in effect sizes for low and mixed SES groups.

Methodology parameters. Parental involvement was a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 6.29$, $p < .05$, $I^2 = 84\%$). Studies involving parents displayed larger effect sizes ($d = 0.71$) than those that did not ($d = 0.38$; see Table 7).

The gender of the dyads significantly moderated achievement outcomes yielding high heterogeneity ($Q_B = 75.72$, $p < .001$, $I^2 = 99\%$). Studies involving same-gender dyads displayed larger effect sizes ($d = 0.85$) than did mixed gender dyads ($d = 0.36$).

Regarding the nature of the test administered for controlling author bias, the outcome measures were categorized into standardized or unstandardized tests. It was found that the nature of the test administered to control for author bias was a significant moderator of achievement outcome and yielded considerable heterogeneity ($Q_B = 18.79$, $p < .001$, $I^2 = 95\%$). Studies in which standardized tests were administered displayed smaller effect sizes ($d = 0.35$) than did those in which unstandardized tests were administered ($d = 0.47$).

The type of reward was also a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 22.60$, $p < .001$, $I^2 = 95\%$). Studies in which tangible items were used as rewards ($d = 0.56$) displayed larger effect sizes than did those involving earning only points as rewards ($d = 0.30$).

The use of teams that competed for rewards significantly moderated the achievement outcome, yielding high heterogeneity ($Q_B = 25.70$, $p < .001$, $I^2 = 96\%$). Studies involving the formation of a competitive team displayed smaller effect sizes ($d = 0.30$) than those that did not ($d = 0.44$). Similarly, the regular formation of new competing teams was a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 16.63$, $p < .01$, $I^2 = 94\%$). Studies involving the formation of new competing teams displayed smaller effect sizes ($d = 0.24$) than those that did not ($d = 0.42$).

Frequency of tutor training was another significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 26.6$, $p < .001$, $I^2 = 96\%$). Studies with 2.5 or fewer sessions of tutoring per week displayed larger effect sizes ($d = 0.46$) than did those with more than 2.5 sessions ($d = 0.26$).

Structured tutoring ($Q_B = 2.39$, *ns*, $I^2 = 58\%$), tutor training ($Q_B = 0.04$, *ns*, $I^2 = 0\%$), length of tutor training sessions ($Q_B = 3.77$, *ns*, $I^2 = 73\%$), fidelity checks ($Q_B = 3.73$, *ns*, $I^2 = 73\%$), mode of assignment of dyad pairs ($Q_B = 1.25$, *ns*, $I^2 = 20\%$), random assignment or experimental versus quasi-experimental design ($Q_B = 0.00$, *ns*, $I^2 = 0\%$), and type of peer tutoring ($Q_B = 0.10$, *ns*, $I^2 = 0\%$) were all nonsignificant moderators of achievement outcome. More specifically, similar effect sizes were found regardless of whether structured tutoring was adopted, tutor training was provided, tutor training sessions exceeded 45 min, fidelity checks were conducted, assignment of dyad pairs was decided by the teacher or by the students themselves, random assignment was used (adoption of experimental or quasi-experimental design), or same-age reciprocal peer, same-age nonreciprocal peer, or cross-age peer tutoring was utilized.

Intervention parameters. The duration of tutoring was a significant moderator of achievement outcome and yielded high heterogeneity ($Q_B = 30.91$, $p < .001$, $I^2 = 97\%$). Studies with 10 or fewer total weeks for the tutoring displayed larger effect sizes ($d = 0.50$) than did those with more than 10 weeks ($d = 0.35$; see Table 8).

Frequency of tutoring sessions ($Q_B = 3.67$, *ns*, $I^2 = 73\%$), length of each tutoring session ($Q_B = 1.88$, *ns*, $I^2 = 47\%$), and total dosage of tutoring ($Q_B = 1.18$, *ns*, $I^2 = 7\%$) were all nonsignificant moderators of achievement outcome. More specifically, similar effect sizes were found regardless of whether the frequency of each tutoring session was more than 2.5 sessions per week, the length of each tutoring session was longer than 30 min, or the total dosage of tutoring was more than 16.25 hr.

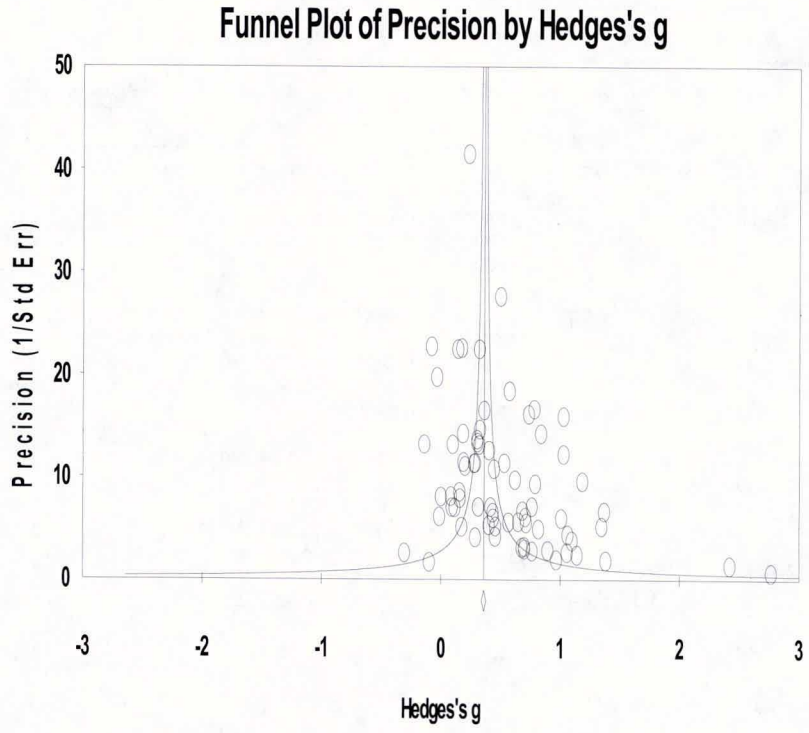

Figure 3. Funnel plot showing effect sizes. Std Err = standard error.

Outcome parameters. Subject content was a significant moderator of achievement outcome when comparing mathematics, reading, and group of other subjects (e.g., physical education or arts), yielding high heterogeneity ($Q_B = 24.33$, $p < .001$, $I^2 = 92\%$). In particular, the use of group of other subjects as subject content displayed the largest effect size ($d = 0.48$), followed by reading ($d = 0.34$), and mathematics ($d = 0.34$). Hence, there was no significant difference in effect sizes between mathematics and reading.

Separating the group of other subjects into different subjects and comparing each of these subjects with mathematics and reading produced high heterogeneity ($Q_B = 157.19$, $p < .001$, $I^2 = 96\%$). In particular, the use of physical education as subject content displayed the largest effect size ($d = 0.90$), followed by arts ($d = 0.82$), science and technology ($d = 0.45$), psychology ($d = 0.39$), reading ($d = 0.34$), mathematics ($d = 0.34$), and language ($d = 0.15$; see Table 9).

The outcome measures on mathematics and reading achievement involve assessment on the subskills of these two subjects. For mathematics, nine studies assessed general arithmetic skills on operation/computation (e.g., subtraction, addition, multiplication), four studies assessed both operation/computation and application of concept (e.g., solve story problems or graphic scenarios) separately, and seven studies assessed conglomerate skills including both operation/computation and application of concept together. It was revealed that the mathematics subskills were a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 21.79$, $p < .001$, $I^2 = 91\%$). In particular, outcome measures on operation/computation displayed the largest effect size ($d = 0.60$), followed by conglomerate skills ($d = 0.30$) and application of concept ($d = 0.24$).

For reading, five studies assessed general comprehension, 16 studies assessed subskills (e.g., word identification or oral reading) separately in addition to general comprehension, six studies assessed subskills (e.g., word identification or oral reading) separately other than general comprehension, and four studies assessed conglomerate skills (e.g., phonics, word identification, sentence and question marks) together. It was revealed that the reading subskills were a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 16.08$, $p < .01$, $I^2 = 75\%$). In particular, outcome measure on maze (replacement of words/phrases) displayed the largest effect size ($d = 0.61$), followed by oral reading ($d = 0.42$), comprehension ($d = 0.32$), word identification ($d = 0.30$), and conglomerate skills ($d = 0.25$).

Test for Moderators of Effects for the Mixed Effects Model

Participant parameters. The educational levels of the participants ($Q_B = 3.69$, *ns*, $I^2 = 19\%$); academic ability of the tutee ($Q_B = 4.09$, *ns*, $I^2 = 27\%$) and tutor ($Q_B = 6.38$, *ns*, $I^2 = 53\%$); proportion of minority of participants ($Q_B = 1.38$, *ns*, $I^2 = 28\%$); and SES of participants ($Q_B = 1.85$, *ns*, $I^2 = 0\%$) were not significant moderators of achievement outcome (see Table 6). Since number of studies in kindergarten group was too small ($n = 2$), it was important to examine whether removal of this group would have significant impact on the result. However, removal of kindergarten group did not affect the result ($Q_B = 2.26$, *ns*, $I^2 = 12\%$), and same order of effect sizes was revealed. Similarly, although number of studies in middle SES was too small ($n = 2$), removal of this SES group did not affect the result ($Q_B = 0.05$, *ns*, $I^2 = 0\%$).

Table 6
Homogeneity Analyses and Mean Effect Size for Possible Participant Parameters as Moderators for Tutee Achievement (k = 72)

Variable	k	Fixed effect—Significant and nonsignificant moderator				Mixed effect—Nonsignificant moderator			
		Q_B	Q_W	d	95% CI	Q_B	Q_W	d	95% CI
Education level of participant[a]		34.74*** (91%)		0.38***	[0.36, 0.41]	3.69 (19%)		0.47***	[0.39, 0.56]
Kindergarten	2		2.86	0.25***	[0.12, 0.37]		0.25	0.21	[−0.23, 0.65]
Elementary	46		337.82***	0.34***	[0.31, 0.38]		57.18	0.47***	[0.37, 0.58]
Secondary school	12		172.51***	0.52***	[0.46, 0.58]		14.79	0.61***	[0.40, 0.81]
College or university	9		36.03***	0.43***	[0.36, 0.51]		4.97	0.38***	[0.16, 0.60]
(mixed level k = 3)									
Academic ability of tutee[a,b]		27.35*** (89%)		0.36***	[0.34, 0.39]	4.09 (27%)		0.46***	[0.38, 0.54]
Low	41		237.56***	0.41***	[0.37, 0.45]		53.89	0.54***	[0.43, 0.65]
Average	13		42.94***	0.29***	[0.23, 0.36]		8.58	0.37***	[0.18, 0.55]
High	11		22.56*	0.44***	[0.36, 0.51]		4.91	0.37***	[0.16, 0.57]
Mixed	10		137.52***	0.26***	[0.20, 0.32]		16.13	0.39***	[0.19, 0.59]
(unspecified group k = 17)									
Academic ability of tutor[a,b]		27.08*** (89%)		0.37***	[0.34, 0.40]	6.38 (53%)		0.47***	[0.39, 0.55]
Low	39		208.47***	0.43***	[0.40, 0.47]		49.21	0.57***	[0.46, 0.68]
Average	13		42.94***	0.29***	[0.23, 0.36]		8.80	0.37***	[0.18, 0.55]
High	13		29.90**	0.39***	[0.32, 0.46]		5.37	0.34***	[0.16, 0.52]
Mixed	9		136.87***	0.27***	[0.20, 0.33]		16.13*	0.42***	[0.21, 0.62]
(unspecified group k = 18)									
Minority of participants[a]		3.89* (74%)		0.37***	[0.34, 0.40]	1.38 (28%)		0.44***	[0.33, 0.55]
≤50%	22		188.66***	0.40***	[0.35, 0.44]		21.99	0.38***	[0.24, 0.53]
>50%	17		135.53***	0.33***	[0.28, 0.38]		18.72	0.52***	[0.35, 0.68]
(unspecified group k = 33)									
SES of participants[a,b]		5.73 (65%)		0.38***	[0.34, 0.41]	1.85 (0%)		0.58***	[0.44, 0.71]
Low	13		63.21***	0.35***	[0.30, 0.40]		15.93	0.59***	[0.39, 0.79]
Middle	2		1.20	0.22*	[0.03, 0.42]		0.26	0.27***	[0.20, 0.74]
Mixed	12		149.82***	0.42***	[0.37, 0.48]		14.74	0.62***	[0.43, 0.82]
(high SES k = 1; unspecified group k = 45)									

Note. A significant Q_B (between-group homogeneity statistic) indicates a significant moderator, whereas a nonsignificant Q_W (within-group homogeneity statistic) shows that the variable can be grouped into homogeneous subgroups. Percentage in parentheses under $Q_B = I^2$ index. k = number of effect sizes, SES = socioeconomic status, CI = confidence interval.

[a] These categories include unspecified groups, mixed group, or subcategories that have degrees of freedom equal to 0. For this reason, the overall value of k for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [b] Some individual samples contributed to effect sizes of more than one group of this moderator variable (see discussion of "Unit of Analysis" in Method for details). For this reason, the overall value of k for the test of this moderator variable is greater than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere.

* $p < .05$. ** $p < .01$. *** $p < .001$.

More specifically, similar effect sizes were found regardless of the educational level of participants, academic ability of the tutee and tutor, proportion of minority of participants, and SES of participants.

Methodology parameters. Structured tutoring was a significant moderator of achievement outcome, yielding high heterogeneity ($Q_B = 5.43$, $p < .05$, $I^2 = 82\%$). Studies adopting structured tutoring displayed larger effect sizes ($d = 0.53$) compared with the adoption of unstructured tutoring ($d = 0.33$; see Table 7).

The type of reward was also a significant moderator of achievement outcome ($Q_B = 6.81$, $p < .01$, $I^2 = 85\%$), yielding high heterogeneity. Studies in which tangible items ($d = 0.70$) were used as rewards displayed larger effect sizes than did those involving just earning points as rewards ($d = 0.35$).

The gender of the dyads significantly moderated achievement outcomes ($Q_B = 13.15$, $p < .001$, $I^2 = 92\%$), yielding high heterogeneity. Studies involving same-gender dyads displayed larger effect sizes ($d = 0.80$) than did those involving mixed gender dyads ($d = 0.41$).

Frequency of tutor training was another significant moderator of achievement outcome ($Q_B = 5.55$, $p < .05$, $I^2 = 82\%$), yielding high heterogeneity. Studies with 2.5 or fewer sessions of tutoring per week displayed larger effect sizes ($d = 0.64$) than did those with more than 2.5 sessions ($d = 0.37$).

However, parental involvement ($Q_B = 1.18$, *ns*, $I^2 = 15\%$); the nature of the test administered to control for author bias ($Q_B = 0.00$, *ns*, $I^2 = 0\%$); the use of teams competing for rewards ($Q_B = 1.83$, *ns*, $I^2 = 45\%$); regular formation of new competing teams ($Q_B = 1.81$, *ns*, $I^2 = 45\%$); tutor training ($Q_B = 3.39$, *ns*, $I^2 = 71\%$); length of tutor training sessions ($Q_B = 0.34$, *ns*, $I^2 = 0\%$); fidelity checks ($Q_B = 0.21$, *ns*, $I^2 = 0\%$); mode of assignment for the tutoring pairs ($Q_B = 0.25$, *ns*, $I^2 = 0\%$); random assignment including experimental and quasi-experimental design ($Q_B = 0.32$, *ns*, $I^2 = 0\%$); and type of peer tutoring ($Q_B = 0.93$, *ns*, $I^2 = 0\%$) were not significant moderators of achievement outcome.

More specifically, similar effect sizes were found regardless of whether parental involvement occurred; assessment controlled for author bias based on the use of standardized or unstandardized

Table 7

Homogeneity Analyses and Mean Effect Size for Possible Methodology Parameters as Moderators for Tutee Achievement (k = 72)

Variable	k	Fixed effect—Significant and nonsignificant moderator				Mixed effect—Significant and nonsignificant moderator			
		Q_B	Q_W	d	95% CI	Q_B	Q_W	d	95% CI
Structured tutoring[a]		2.39 (58%)		0.30***	[0.36, 0.41]	5.43* (82%)		0.47***	[0.39, 0.55]
Yes	53		500.40***	0.40***	[0.37, 0.43]		69.51	0.53***	[0.44, 0.63]
No	23		90.91***	0.35***	[0.30, 0.40]		15.84	0.33***	[0.18, 0.47]
(mixed type k = 4)									
Parent involvement[a]		6.29* (84%)		0.39***	[0.36, 0.41]	1.18 (15%)		0.48***	[0.40, 0.56]
Yes	3		3.47	0.71***	[0.46, 0.97]		1.09	0.71**	[0.28, 1.14]
No	70		582.22***	0.38***	[0.36, 0.41]		87.7	0.47***	[0.39, 0.55]
(mixed type k = 1)									
Same-gender dyad		75.72*** (99%)		0.39***	[0.36, 0.41]	13.15*** (92%)		0.47***	[0.39, 0.55]
Same gender	14		18.71	0.85***	[0.75, 0.96]		6.79	0.80***	[0.61, 1.00]
Mixed gender	58		495.03***	0.36***	[0.33, 0.39]		73.09	0.41***	[0.33, 0.49]
Control for author bias[a,b] (standardized test used)		18.79*** (95%)		0.39***	[0.36, 0.41]	0.00 (0%)		0.47***	[0.39, 0.55]
Yes	41		303.92***	0.35***	[0.32, 0.38]		50.18	0.47***	[0.37, 0.58]
No	32		264.46***	0.47***	[0.42, 0.51]		37.47	0.47***	[0.34, 0.59]
(mixed type k = 3)									
Reward type[b]		22.06*** (95%)		0.34***	[0.30, 0.38]	6.81** (85%)		0.46***	[0.34, 0.57]
Earning points	12		109.16***	0.30***	[0.26, 0.34]		14.52	0.35***	[0.21, 0.48]
Tangible items	9		41.91	0.56***	[0.46, 0.66]		12.46	0.70***	[0.48, 0.91]
(no reward specified k = 50; mixed type k = 1)									
Formation of competing team for group reward		25.70*** (96%)		0.39***	[0.36, 0.41]	1.83 (45%)		0.47***	[0.39, 0.55]
Yes	12		108.60***	0.30***	[0.26, 0.34]		10.56	0.37***	[0.19, 0.54]
No	60		455.16***	0.44***	[0.40, 0.47]		73.35	0.50***	[0.41, 0.59]
Fidelity check		3.73 (73%)		0.39***	[0.36, 0.41]	0.21 (0%)		0.48***	[0.39, 0.56]
Yes	39		283.41***	0.37***	[0.34, 0.40]		36.99	0.49***	[0.38, 0.61]
No	33		302.32***	0.42***	[0.38, 0.46]		40.40	0.45***	[0.32, 0.58]
Assignment of pairing		1.25 (20%)		0.39***	[0.36, 0.41]	0.25 (0%)		0.47***	[0.39, 0.55]
Voluntary basis	2		0.07	0.61**	[0.22, 1.01]		0.04	0.62*	[0.05, 1.19]
Was assigned	70		588.13***	0.38***	[0.36, 0.41]		85.27	0.47***	[0.39, 0.55]
Regular formation of new competing team for team reward[c]		16.63** (94%)		0.31***	[0.26, 0.34]	1.81 (45%)		0.35***	[0.22, 0.49]
Yes	7		16.42*	0.24***	[0.19, 0.29]		3.14	0.27**	[0.10, 0.45]
No	5		75.55***	0.42***	[0.35, 0.49]		11.73*	0.47***	[0.25, 0.68]
(no team k = 60)									
Frequency of tutor training[b] (number of session)		26.60*** (96%)		0.33***	[0.29, 0.37]	5.55* (82%)		0.49***	[0.37, 0.60]
≤2.5 sessions per week	17		179.90***	0.46***	[0.40, 0.53]		33.72*	0.64***	[0.47, 0.81]
>2.5 sessions per week (median = 2.5 sessions per week; unspecified group k = 37)	18		54.41***	0.26***	[0.22, 0.31]		13.01	0.37***	[0.21, 0.52]
Tutor training[a]		0.04 (0%)		0.39***	[0.36, 0.41]	3.39 (71%)		0.51***	[0.43, 0.60]
Yes	55		584.45***	0.39***	[0.36, 0.42]		114.02***	0.57***	[0.46, 0.67]
No	23		163.94***	0.38***	[0.34, 0.43]		15.81	0.38***	[0.22, 0.55]
(mixed mode k = 6)									
Length of tutor training[b] (minutes per session)		3.77 (73%)		0.33***	[0.29, 0.36]	0.34 (0%)		0.48***	[0.36, 0.59]
≤45 min per session	18		116.52***	0.29***	[0.24, 0.34]		25.21	0.51***	[0.35, 0.67]
>45 min per session	15		126.67***	0.36***	[0.31, 0.41]		19.41	0.44***	[0.27, 0.61]

(*table continues*)

Table 7 (*continued*)

Variable	*k*	Fixed effect—Significant and nonsignificant moderator				Mixed effect—Significant and nonsignificant moderator			
		Q_B	Q_W	*d*	95% CI	Q_B	Q_W	*d*	95% CI
(median = 45 min per session; unspecified group *k* = 39)									
Random assignment[a]		0.00 (0%)		0.39***	[0.36, 0.41]	0.32 (0%)		0.47***	[0.38, 0.55]
Yes (experimental)	53		442.20***	0.39***	[0.36, 0.41]		59.15	0.48***	[0.38, 0.58]
No (quasi-experimental)	20		227.43***	0.38***	[0.33, 0.44]		27.62	0.43***	[0.26, 0.59]
(mixed mode *k* = 1)									
Type of peer tutoring[a,b]		0.10 (0%)		0.39***	[0.36, 0.41]	0.93 (0%)		0.47***	[0.39, 0.56]
Same-age reciprocal peer tutoring	42		264.25***	0.38***	[0.35, 0.42]		40.32	0.50***	[0.39, 0.62]
Same-age nonreciprocal peer tutoring	15		189.57***	0.38***	[0.33, 0.44]		18.32	0.40***	[0.22, 0.59]
Cross-age peer tutoring (mixed type *k* = 5)	16		88.12***	0.39***	[0.33, 0.46]		20.70	0.45***	[0.26, 0.65]

Note. A significant Q_B (between-group homogeneity statistic) indicates a significant moderator, whereas a nonsignificant Q_W (within-group homogeneity statistic) shows that the variable can be grouped into homogeneous subgroups. Percentage in parentheses under $Q_B = I^2$ index. *k* = number of effect sizes, CI = confidence interval.

[a] Some individual samples contributed effect sizes to more than one group of this moderator variable (see discussion of "Unit of Analysis" in Method for details). For this reason, the overall value of *k* for the test of this moderator variable is greater than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [b] These categories include unspecified groups, mixed group or subcategories which have degrees of freedom equal to 0. For this reason, the overall value of *k* for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [c] These categories exclude those without forming competing team, which are not applicable for the analysis. For this reason, the overall value of *k* for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

tests; team formation to compete for group rewards or regular formation of new competing teams occurred; tutor training was provided; tutor training sessions exceeded 45 min; fidelity checks occurred; tutoring pairs were assigned or selected by the students; random assignment occurred (adoption of experimental or quasi-experimental design); or the peer tutoring was same-age reciprocal, same-age nonreciprocal, or cross-age in nature.

Intervention parameters. The duration of tutoring (Q_B = 1.99, *ns*, I^2 = 50%), frequency of tutoring sessions (Q_B = 0.00, I^2 = 0%), length of each tutoring session (Q_B = 0.13, I^2 = 0%), and total dosage of tutoring (Q_B = 0.26, I^2 = 0%) were not significant moderators of achievement. More specifically, similar effect sizes were found regardless of whether the frequency of each tutoring session was more than 2.5 sessions per week, the length of each tutoring session was longer than 30 min, the total dosage of tutoring was more than 16.25 hr, or the tutoring lasted for more than 10 weeks (see Table 8).

Outcome parameters. Subject content was not a significant moderator of achievement outcome when comparing mathematics, reading, and group of other subjects (e.g., physical education or arts; Q_B = 1.45, *ns*, I^2 = 0%; see Table 9). Similarly, when separating the group of other subjects into different subjects, subject content was also not a significant moderator of achievement outcome (Q_B = 13.88, *ns*, I^2 = 50%). More specifically, similar effect sizes were found regardless of the subject content tutored.

Mathematics subskills were a significant moderator of achievement outcome, yielding medium heterogeneity (Q_B = 6.06, $p < .05$, I^2 = 67%). In particular, outcome measure on operation/computation displayed the largest effect size (*d* = 0.69), followed by conglomerate skills (*d* = 0.43) and application of concept (*d* = 0.24).

Reading subskills were not a significant moderator of achievement outcome (Q_B = 1.25, *ns*, I^2 = 20%). More specifically, similar effect sizes were found regardless of what type of subskills was assessed.

Test for Interaction Between Moderators of Effects

Fixed effects model. There was significant interaction between educational levels of the participants and minority of participants, yielding high heterogeneity (Q_B = 20.25, $p < .001$, I^2 = 85%; see Table 10). Studies involving secondary school students and with 50% or fewer participants from minority groups displayed larger effect sizes (*d* = 0.68) than those involving secondary school students and with more than 50% minority participants (*d* = 0.30). However, studies involving elementary school students and with 50% or fewer participants from minority groups displayed only a slight larger effect sizes (*d* = 0.41) than those involving elementary school students and with more than 50% minority participants (*d* = 0.39).

Moreover, there was significant interaction between the subject matter and the test administered for controlling for author bias (using standardized test), yielding high heterogeneity (Q_B = 106.03, $p < .001$, I^2 = 95%). In particular, when adopting standardized test, the use of mathematics as subject content displayed the largest effect size (*d* = 0.45), followed by reading (*d* = 0.34), and group of other subjects (*d* = 0.15). In contrast, when adopting unstandardized test, the use of group of other subjects as subject

Table 8
Homogeneity Analyses and Mean Effect Size for Possible Intervention Parameters as Moderators for Tutee Achievement (k = 72)

Variable	k	Fixed effect—Significant and nonsignificant moderator				Mixed effect—Nonsignificant moderator			
		Q_B	Q_W	d	95% CI	Q_B	Q_W	d	95% CI
Duration of tutoring[a] (no. of weeks)		30.91*** (97%)		0.40***	[0.37, 0.42]	1.99 (50%)		0.48***	[0.39, 0.56]
≤10 weeks	32		261.77***	0.50***	[0.46, 0.55]		46.37*	0.54***	[0.42, 0.67]
>10 weeks	28		221.21***	0.35***	[0.32, 0.38]		22.96	0.42***	[0.30, 0.53]
(median = 10 weeks; unspecified group k = 12)									
Frequency of each tutoring session[a] (no. of session per week)		3.67 (73%)		0.39***	[0.37, 0.42]	0.00 (0%)		0.50***	[0.42, 0.58]
≤ 2.5 sessions per week	29		206.87***	0.42***	[0.38, 0.46]		38.50	0.50***	[0.39, 0.61]
>2.5 sessions per week	25		127.22***	0.37***	[0.33, 0.41]		34.00	0.50***	[0.38, 0.63]
(median = 2.5 sessions per week; unspecified group k = 18)									
Length of each tutoring session[a] (minutes per session)		1.88 (47%)		0.39***	[0.36, 0.42]	0.13 (0%)		0.47***	[0.38, 0.57]
≤30 min per session	29		395.40***	0.37***	[0.34, 0.41]		51.85**	0.46***	[0.33, 0.58]
>30 min per session	25		101.14***	0.41***	[0.37, 0.45]		14.55	0.49***	[0.35, 0.63]
(median = 30 min per session; unspecified group k = 18)									
Total dosage of tutoring[a] (number of hour)		1.18 (7%)		0.41***	[0.38, 0.44]	0.26 (0%)		0.50***	[0.41, 0.59]
≤16.25 hr	22		106.29***	0.44***	[0.38, 0.49]		37.27*	0.47***	[0.34, 0.61]
>16.25 hr	21		160.37***	0.40***	[0.37, 0.44]		20.72	0.52***	[0.40, 0.64]
(median = 16.25 hr; unspecified group k = 29)									

Note. A significant Q_B (between-group homogeneity statistic) indicates a significant moderator, whereas a nonsignificant Q_W (within-group homogeneity statistic) shows that the variable can be grouped into homogeneous subgroups. Percentage in parentheses under $Q_B = I^2$ index. k = number of effect sizes, CI = confidence interval.

[a] These categories include unspecified groups, mixed group, or subcategories which have degrees of freedom equal to 0. For this reason, the overall value of k for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere.

* $p < .05$. ** $p < .01$. *** $p < .001$.

content displayed the largest effect size ($d = 0.58$), followed by reading ($d = 0.36$) and mathematics ($d = 0.07$).

Mixed effects model. There was no significant interaction between educational levels of the participants and minority of participants ($Q_B = 3.51$, *ns*, $I^2 = 15\%$) or between subject matter and test administered for controlling for author bias (using standardized test; $Q_B = 6.87$, *ns*, $I^2 = 27\%$).

Discussion

The meta-analysis conducted in the present study advances the knowledge of the effects of peer tutoring on achievement. The present updated meta-analysis offers stronger support than previous meta-analyses for the overall effectiveness of peer tutoring on academic achievement by addressing the limitations of previous meta-analytic research, including studies that examined a greater range of subject content and participants and by adopting current methodological advances in meta-analysis. Moreover, the present investigation adds to the understanding of how the intervention features moderate the effects of peer tutoring, including those features that were not evaluated in previous meta-analyses.

Overall Impact of Peer Tutoring on Achievement

The meta-analysis conducted in the present investigation provides clear evidence that peer tutoring has a positive effect on the academic achievements of tutees. After imputing the missing values by adopting the trim-and-fill method, the weighted mean effect size was 0.26 (95% CI [0.24, 0.28]) for the fixed effects model, whereas it was 0.37 (95% CI [0.29, 0.45]) for the mixed effects model. The results indicated that the posttest scores were greater in the treatment groups than in the control groups.

Moderators of the Effects of Peer Tutoring on Achievement

Fixed effects model. The present investigation identified specific features of the participants, including the educational levels of the participants, the academic ability levels of both tutees and tutors, and the proportions of minority of students, that were all significant moderators of effect size.

Regarding the educational levels of the participants, the studies of participants from secondary schools displayed larger effect sizes than did those of students at other educational levels, followed by college- or university-level, elementary school, and kindergarten participants. Because previous meta-analyses have not compared the effects of educational levels, the present investigation adds a new understanding of how educational level moderates the effects of peer tutoring. However, caution should be taken when including kindergarten participants for interpretation since the number of studies was too small ($n = 2$), which would reduce the statistical power of the analyses.

For the academic ability of the tutee, studies of tutees with high ability displayed larger effect sizes than did those of other aca-

Table 9
Homogeneity Analyses and Mean Effect Size for Possible Outcome Parameters as Moderators for Tutee Achievement (k = 72)

Variable	k	Fixed effect—Significant moderator Q_B	Q_W	d	95% CI	Mixed effect—Significant and nonsignificant moderator Q_B	Q_W	d	95% CI
Target subject matter on achievement[a,b]		24.33*** (92%)		0.39***	[0.36, 0.41]	1.45 (0%)		0.45***	[0.38, 0.52]
Mathematics	20		125.58***	0.34***	[0.27, 0.41]		24.93	0.53***	[0.38, 0.67]
Reading	31		215.64**	0.34***	[0.31, 0.38]		36.15	0.42***	[0.31, 0.53]
Other subjects (mixed subject k = 4)	27		228.31***	0.48***	[0.43, 0.53]		30.44	0.42***	[0.29, 0.55]
Target subject matter on achievement (with different separate subjects in other subject subgroup)[a,b]		157.19*** (96%)		0.39***	[0.36, 0.41]	13.88 (50%)		0.45***	[0.38, 0.52]
Mathematics	20		125.58***	0.34***	[0.27, 0.41]		28.81	0.53***	[0.38, 0.67]
Reading	31		215.64**	0.34***	[0.31, 0.38]		40.79	0.42***	[0.31, 0.53]
Language	6		14.12*	0.15**	[0.05, 0.25]		6.01	0.21	[−0.06, 0.48]
Science and technology	6		18.56**	0.45***	[0.37, 0.53]		2.64	0.35**	[0.11, 0.60]
Physical education	4		7.12	0.90***	[0.72, 1.07]		2.30	0.84***	[0.51, 1.18]
Arts	4		29.72***	0.82***	[0.73, 0.91]		3.34	0.67***	[0.39, 0.95]
Psychology	4		14.29**	0.39***	[0.21, 0.57]		4.79	0.34*	[0.01, 0.68]
Miscellaneous subject (mixed subject k = 4)	3		2.22	0.19	[−0.01, 0.39]		1.62	0.26	[−0.14, 0.66]
Subskills on mathematics achievement[b]		21.79*** (91%)		0.42***	[0.35, 0.49]	6.06* (67%)		0.53***	[0.38, 0.67]
Operation/computation	13		54.94***	0.60***	[0.50, 0.70]		14.16	0.69***	[0.49, 0.89]
Application of concept	4		5.01	0.24***	[0.11, 0.38]		0.98	0.24	[−0.08, 0.57]
Conglomerate skills	7		16.97**	0.30***	[0.17, 0.43]		4.36	0.43***	[0.16, 0.70]
Subskills on reading achievement[b]		16.08** (75%)		0.36***	[0.30, 0.42]	1.25 (20%)		0.39***	[0.30, 0.49]
Comprehension	21		29.83	0.32***	[0.22, 0.43]		16.42	0.42***	[0.26, 0.58]
Word identification	10		12.12	0.30***	[0.16, 0.45]		7.94	0.31**	[0.09, 0.54]
Oral reading	18		37.83**	0.42***	[0.28, 0.57]		24.34	0.43***	[0.24, 0.61]
Maze (replace missing words or phrases)	7		19.88**	0.61***	[0.46, 0.77]		11.22	0.44**	[0.17, 0.72]
Conglomerate skills	5		7.59	0.25***	[0.14, 0.37]		5.12	0.30*	[0.02, 0.58]

Note. A significant Q_B (between-group homogeneity statistic) indicates a significant moderator, whereas a nonsignificant Q_W (within-group homogeneity statistic) shows that the variable can be grouped into homogeneous subgroups. Percentage in parentheses under $Q_B = I^2$ index. k = number of effect sizes, CI = confidence interval.

[a] These categories include unspecified groups, mixed group, or subcategories which have degrees of freedom equal to 0. For this reason, the overall value of k for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [b] Some individual samples contributed effect sizes to more than one group of this moderator variable (see discussion of "Unit of Analysis" in Method for details). For this reason, the overall value of k for the test of this moderator variable is greater than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere.

* $p < .05$. ** $p < .01$. *** $p < .001$.

demic ability levels, followed by low, average, and mixed ability levels. For the academic ability of the tutor, the studies of tutors of low ability displayed larger effect sizes than did those of other academic ability levels, followed by high, average, and mixed ability levels. Cohen et al. (1982) found that there was a greater, although not significant, effect on size for tutees with low academic ability levels (unweighted ES = 0.42) than for those with middle academic ability levels (unweighted ES = 0.33). A review conducted by Robinson et al. (2005) noted that there is evidence that low-achieving students benefit from tutoring (e.g., Ginsburg-Block & Fantuzzo, 1997; Simmons, Fuchs, Fuchs, Hodge, & Mathes, 1994; Sprinthall & Scott, 1989). Similarly, studies of tutors with low ability levels have displayed larger effect sizes than those of students at other academic ability levels. Hence, low-performing tutees and tutors both benefit from tutoring (Polirstok & Greer, 1986). Therefore, the findings of the present study are consistent with previous studies to a certain extent. However, previous meta-analytic studies have not demonstrated that tutees and tutors of high ability produce greater effects than do those of other academic ability levels. Hence, the present study increases understanding of how the academic ability of both the tutee and tutor moderates the effects of peer tutoring on achievement.

For minority participants, studies with 50% or fewer minority participants exhibited larger effects than studies with more than 50% minority participants. However, Rohrbeck et al. (2003) reported a significantly greater effect size of achievement in studies with more than 50% ethnic minority students (weighted ES = 0.51) than in studies with fewer minority students (weighted ES = 0.23). These inconsistent findings may be due to differences in the intervention modes, such as peer tutoring in the present investigation, whereas other interventions (such as cooperative learning) were included in the study conducted by Rohrbeck et al. (2003). Additionally, the present study included participants of different educational levels (ranging from kindergarten to college/university), whereas only elementary students were included in the study conducted by Rohrbeck et al. (2003). Nevertheless, the present

Table 10
Homogeneity Analyses and Mean Effect Size for Interaction Between Some Moderators for Tutee Achievement (k = 72)

Variable	k	Fixed effect—Significant moderator				Mixed effect—Nonsignificant moderator			
		Q_B	Q_W	d	95% CI	Q_B	Q_W	d	95% CI
Education Level of Participant × Minority[a]		20.25*** (85%)		0.40***	[0.36, 0.43]	3.51 (15%)		0.49***	[0.36, 0.63]
Elementary–minority ≤50%	13		128.13***	0.41***	[0.36, 0.46]		8.75	0.36***	[0.16, 0.57]
Elementary–minority >50%	11		90.25***	0.39***	[0.30, 0.48]		11.81	0.61***	[0.37, 0.85]
Secondary–minority ≤50%	5		20.70***	0.68***	[0.53, 0.84]		5.32	0.68***	[0.31, 1.04]
Secondary–minority >50%	3		33.06***	0.30***	[0.23, 0.37]		1.28	0.44***	[0.33, 0.84]
Target Subject Matter × Control for Author Bias[a,b,c]		106.03*** (95%)		0.38***	[0.36, 0.41]	6.87 (27%)		0.45***	[0.37, 0.52]
Mathematics–standardized test	16		83.37***	0.45***	[0.37, 0.53]		20.27	0.56***	[0.39, 0.73]
Reading–standardized test	27		207.21**	0.34***	[0.31, 0.38]		35.07	0.40***	[0.28, 0.52]
Other subject–standardized test	6		10.40	0.15**	[0.05, 0.25]		2.79	0.18	[−0.09, 0.44]
Mathematics–unstandardized test	5		18.14**	0.07	[−0.05, 0.20]		5.89	0.42**	[0.08, 0.76]
Reading–unstandardized test	5		7.33	0.36***	[0.22, 0.50]		4.80	0.54**	[0.13, 0.94]
Other subject–unstandardized test	21		159.12***	0.58***	[0.52, 0.63]		24.90	0.50***	[0.36, 0.64]

Note. A significant Q_B (between-group homogeneity statistic) indicates a significant moderator, whereas a nonsignificant Q_W (within-group homogeneity statistic) shows that the variable can be grouped into homogeneous subgroups. Percentage in parentheses under $Q_B = I^2$ index. k = number of effect sizes, CI = confidence interval.

[a] These categories include unspecified groups, mixed group or subcategories which have degrees of freedom equal to 0. For this reason, the overall value of k for the test of this moderator variable is smaller than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [b] Some individual samples contributed effect sizes to more than one group of this moderator variable (see discussion of "Unit of Analysis" in methodology for details). For this reason, the overall value of k for the test of this moderator variable is greater than the total number of independent samples (i.e., 72). This also explains why the overall mean effect size for these variables is different from the mean of 0.39 in fixed effect model and 0.47 in mixed effect model reported elsewhere. [c] Standardized test used.

** $p < .01$. *** $p < .001$.

investigation increases the understanding of how minority status moderates the effects of peer tutoring by including studies of different educational levels and focusing on peer tutoring.

Also, since there was significant interaction between educational levels of the participants and minority of participants, studies involving secondary school students and with 50% or fewer participants from minority groups would display larger effect sizes than those involving elementary school students and with 50% or fewer participants from minority groups.

Regarding the methodology parameters, parental involvement, the provision of fewer tutor training sessions, the use of tangible rewards, the use of unstandardized tests, the adoption of same-gender dyads, and no formation of competing teams for group rewards were all significant moderators of effect size. Studies utilizing parental involvement displayed greater effect sizes than did those without parent involvement. Because previous meta-analyses have not compared the effects of parental involvement, the present study provides preliminary evidence to suggest that home–school cooperation in peer tutoring is important for promoting the academic achievement of students. However, caution should be taken since the number of studies involving parents is too small ($n = 2$), which would reduce the statistical power of the analyses.

Regarding the parameters of the tutor training, fewer weekly training sessions of tutor training produced greater effect sizes than did more frequent training sessions. Because previous meta-analyses have not examined the effects of these parameters, the present investigation adds new understanding concerning how the number of weekly training sessions moderates the effects of peer tutoring.

The studies where tutees were given tangible rewards displayed larger effect sizes than did those in which the tutees received points as rewards. Because previous meta-analyses have not investigated the effects of these types of rewards, the present investigation adds to the understanding of how such rewards moderate the effects of peer tutoring.

Regarding the test for assessing the outcomes, using unstandardized tests to control for author bias displayed larger effect sizes than using standardized tests. Cohen et al. (1982) reported that there was a greater effect size for studies using unstandardized tests (unweighted ES = 0.84) compared with studies using standardized tests (unweighted ES = 0.27). Cook et al. (1985) also reported that there was a greater effect size for studies using unstandardized tests (unweighted ES = 0.89) compared with studies using standardized tests (unweighted ES = 0.45). However, the results of unstandardized tests need to be interpreted cautiously because they can be constructed (by design or inadvertently) to measure characteristics that are idiosyncratic to an intervention and are not reflective of more general achievement measures.

Moreover, there was significant interaction between subject matter and test administered for controlling for author bias (using standardized test). When adopting standardized test, the use of mathematics and reading as subject content displayed larger effect size than the group of other subjects. On the other hand, when adopting unstandardized test, the use of the group of other subjects and reading as subject content displayed larger effect size than mathematics.

Regarding the theoretically based program parameters, it was hypothesized that both cross-age peer tutoring and same-age reciprocal peer tutoring would produce greater effects than same-age nonreciprocal peer tutoring, according to role theory. However, it was revealed that same-age reciprocal and nonreciprocal peer tutoring and cross-age peer tutoring displayed significantly similar effect sizes. Hence, role theory was not supported, and concern

over the unequal status between tutee and tutor was not established. Nevertheless, this finding contributes to research on peer tutoring because no previous meta-analysis of peer tutoring has simultaneously compared the effects of these three types of tutoring.

Regarding the gender composition of the tutoring dyads, it was hypothesized that same-sex dyads would produce greater effects than would mixed-sex dyads, according to role theory. Consistent with predictions based on role theory, the studies of same-gender dyads displayed larger effect sizes than did those of mixed gender dyads. Hence, the negative impact of gender stereotype and gender role based on role theory was supported. Rohrbeck et al. (2003) found that studies with same-gender dyads (weighted ES = 0.63) displayed larger effect sizes than did those with mixed gender dyads (weighted ES = 0.30).

In addition, the present investigation explored whether forming teams to compete for rewards, as is commonly found in cooperative learning, produces greater effect sizes than peer tutoring without competing teams, consistent with interdependent group reward contingencies. Contrary to prior predictions regarding the interdependent group reward contingency approach, the formation of teams competing for rewards did not display a greater effect size than those studies that did not involve the formation of competing teams. This finding suggests that the adoption of interdependent group reward contingencies could not produce a greater positive gain in achievement as revealed in cooperative learning. In addition, the formation of new competing teams regularly displayed smaller effect sizes than did studies without the regular formation of new competing teams. This finding indicates that although the regular formation of new competing teams has the advantages of equity and assurance of equal chances to win rewards, it did not enhance the effectiveness of peer tutoring. This finding contributes to new understanding on peer tutoring since it clarifies that it is not more effective for peer tutoring to incorporate elements of other peer-assisted interventions, such as cooperative learning.

Regarding the intervention parameter, consistent with previous meta-analyses, a shorter duration of the entire tutoring period displayed larger effect sizes than longer duration. In an earlier meta-analysis, Cohen et al. (1982) reported that a shorter duration of tutoring produced a greater effect on tutees than did interventions that had a longer duration. The effect sizes for durations ranging from 0 to 4 weeks, 5 to 18 weeks, and 19 to 36 weeks were 0.95, 0.42, and 0.16, respectively.

Regarding the subject content in the outcome parameter, two commonly tutored subjects (i.e., mathematics and reading) displayed similar effect sizes while the group of other subjects (e.g., physical education and arts) displayed greater effect than mathematics and reading. Rohrbeck et al. (2003) also found no significant difference between mathematics (weighted ES = 0.27) and reading (weighted ES = 0.26). When separating the group of other subjects into different subjects, the use of physical education as subject content displayed the largest effect sizes, followed by arts, science and technology, psychology, reading, mathematics, and language. However, previous meta-analyses have reported mixed results regarding the relative effects of peer tutoring on mathematics, reading, and other subjects. Cohen et al. (1982) reported a greater effect size for mathematics (unweighted ES = 0.60) than for reading (unweighted ES = 0.29) and other subjects (unweighted ES = 0.30). In contrast, Cook et al. (1985) reported a smaller effect size for mathematics (unweighted ES = 0.85) than for other subjects, such as language (unweighted ES = 1.13). The mixed findings of previous meta-analyses may be related to the methodologies and participants involved, as mentioned previously. Because previous meta-analyses have not compared the effects of more different subjects, the present investigation adds to the understanding of how various subjects moderate the effects of peer tutoring, although the impact of peer tutoring on specific types of subject content needs to be elucidated by future research. However, caution should be taken since the number of studies for each subject in the group of other subjects was too small (*n*s = 3–6), which would reduce the statistical power of the analyses.

Regarding the assessment of the subskills of mathematics achievement, the outcome measure on operation/computation displayed the largest effect size, followed by conglomerate skills and application of concept. For reading achievement, the outcome measure on maze (replacement of words/phrases) displayed the largest effect size, followed by oral reading, comprehension, word identification, and conglomerate skills. Because previous meta-analyses did not compare the effects of assessment of mathematics and reading subskills, the present investigation adds to the understanding of how assessment of subskills of these two commonly tutored subjects moderates the effects of peer tutoring. However, caution should be taken when including assessing application of mathematics concept ($n = 4$) and conglomerate reading skills ($n = 5$) for analyses since the number of studies for these two subgroups is too small, which would reduce the statistical power of the analyses.

Mixed effects model. It was revealed that only tutor training sessions, type of reward, dyad gender and provision of structure for tutoring, and assessment of mathematics subskills were significant moderators of effect size. As in the fixed effects model, fewer weekly training sessions produced larger effect sizes than did more frequent training sessions, the use of tangible rewards produced larger effects sizes than did using points earned as rewards, the adoption of same-gender dyads produced larger effect sizes than did the adoption of mixed-gender dyads, and the outcome measure on operation/computation displayed the larger effect sizes than conglomerate skills and application of concept.

Unlike the findings observed in the fixed effects model, the adoption of structured tutoring produced larger effect sizes than did the adoption of unstructured tutoring under the mixed effects model. In an earlier meta-analysis, Cohen et al. (1982) reported that there was a greater effect size for structured tutoring (unweighted ES = 0.51) than for unstructured tutoring (unweighted ES = 0.26). Hence, the findings of the present study are consistent with those of previous meta-analyses.

Inferences from fixed and mixed effects models. As discussed previously, different inferences would be drawn from the fixed and mixed effects models because the variability in effect size observed in the former comes from participant-level sampling error, whereas in the latter, study-level variance is a further source of errors in addition to the participant-level sampling error. Hence, similar results would be obtained if the present studies were replicated in an identical manner but different participants were drawn from the same population in the fixed effects model. For mixed effects model, similar results would be found if the present studies were replicated in an identical manner even with the adoption of different study characteristics such as methodology

and intervention features and engaging different participants drawn from the same population (O'Mara et al., 2006). In the present investigation, the significant moderators found in the fixed effects model included participants' educational levels, academic abilities of both tutee and tutor, minority of participants, parental involvement, gender of the dyads, nature of the test administered to control for author bias, type of reward, use of competing teams for rewards, regular formation of new competing teams, frequency of tutor training, duration of tutoring, subject content, and assessment of mathematics and reading subskills. It suggests that these moderators could be obtained if different participants were selected from same population included in the present study.

For mixed effects models, only the adoption of structured tutoring, same-gender dyads, tangible rewards, fewer training sessions, assessment of mathematics subskills were significant moderators. It indicates that these moderators could be found on all peer tutoring interventions beyond the studies included in the present investigation.

Implications

Theory-driven parameters. The present investigation provides evidence for the applicability of role theory and the concept of interdependent group contingencies by evaluating the moderating effects of certain theory-driven program features that were not tested in previous meta-analytic studies. It was found that some parameters (e.g., the gender composition of the tutoring dyads) were supported and that others (e.g., the type of peer tutoring and interdependent group reward contingencies) were rejected. Hence, the present meta-analysis suggests other theoretically based program parameters could be examined in similar manner in the future.

School practice and teacher training. From a practical perspective, this updated meta-analysis provides a strong empirical basis for educational practitioners to design and implement peer tutoring confidently to promote academic achievement because it helps identify the program features that can produce optimal effectiveness for peer tutoring.

In essence, based on the findings from both fixed and mixed effects models, the crucial determinants for the optimal effectiveness of peer tutoring on academic achievement include the following:

1. Regarding the participant parameters, studies involving participants in secondary school, followed by college- or university-level, elementary school, and kindergarten participants; tutees of high ability, followed by low, average, and mixed ability levels; tutors of low ability, followed by high, average, and mixed ability levels; and fewer than 50% minority participants displayed the largest effect sizes.

2. Regarding the methodology parameters, studies adopting structured tutoring, parental involvement, unstandardized testing to control for author bias, tutor training with less frequent weekly training sessions, rewards in the form of tangible items, same-gender dyads, no formation of competing teams for group reward, and no regular formation of new teams displayed larger effect sizes.

3. Studies with shorter duration of tutoring produced larger effect sizes.

4. Regarding the outcome parameters, comparing the effect sizes of the two commonly tutored subjects (reading and mathematics), the effect size of mathematics was not significantly different from reading. In comparison of the group of subjects (e.g., physical education and arts) other than these two commonly tutored subjects (reading and mathematics), the group of other subjects displayed larger effect size than mathematics and reading. When separating the group of other subjects into different subjects, the use of physical education as subject content displayed the largest effect sizes, followed by arts, science and technology, psychology, reading, mathematics, and language. Regarding the assessment of the subskills of mathematics achievement, outcome measure on operation/computation displayed the largest effect size, followed by conglomerate skills and application of concept. For reading achievement, outcome measure of maze (replacement of words/phrases) displayed the largest effect size, followed by oral reading, comprehension, word identification, and conglomerate skills.

Hence, a preliminary empirical model of the crucial determinants of best practices for peer tutoring on achievement was proposed (see Figure 4). This model consists of certain components: preintervention, intervention, postintervention, and motivational component in the entire peer-tutoring intervention process. In the preintervention phase, the model is concerned with participant selection, screening, and training. In the intervention phase, the model is concerned with the intervention format and subject content. During the intervention, motivational component plays a crucial role in engaging the participants. In the postintervention phase, outcome assessment is the key focus.

Specifically, the following practices can promote higher academic achievement of tutee. When designing peer tutoring interventions during the preintervention phase, selecting participants from secondary school will be most effective, followed by post-secondary, elementary school, and kindergarten participants. Additionally, selecting tutees that have high ability will be more effective, followed by those with low, average and mixed ability levels. Similarly, selecting tutors with low ability will be better, followed by those with high, average and mixed ability levels. Moreover, selecting a relatively small proportion ($\leq$50%) of participants from ethnic minority backgrounds would also be more effective. However, since there was significant interaction between educational levels of the participants and minority of participants, selection of secondary school students and with 50% or fewer participants from minority groups would display larger effect sizes than those involving elementary school students and with 50% or fewer participants from minority groups. Regarding the tutor training, the provision of less frequent training sessions would be more beneficial.

During the intervention phase, when assigning participants into pairs, it is important to put tutees and tutors of the same sex into pairs. Moreover, regarding the intervention format, it is recommended both to provide a structure for the tutoring that the tutee and tutor can follow and that a shorter duration of intervention should be adopted. In choosing the core tutored subject, both mathematics and reading are equally effective. When choosing a subject other than the two core tutored subjects (mathematics and reading), physical education is the most effective subject, followed by arts, science and technology, psychology, and language.

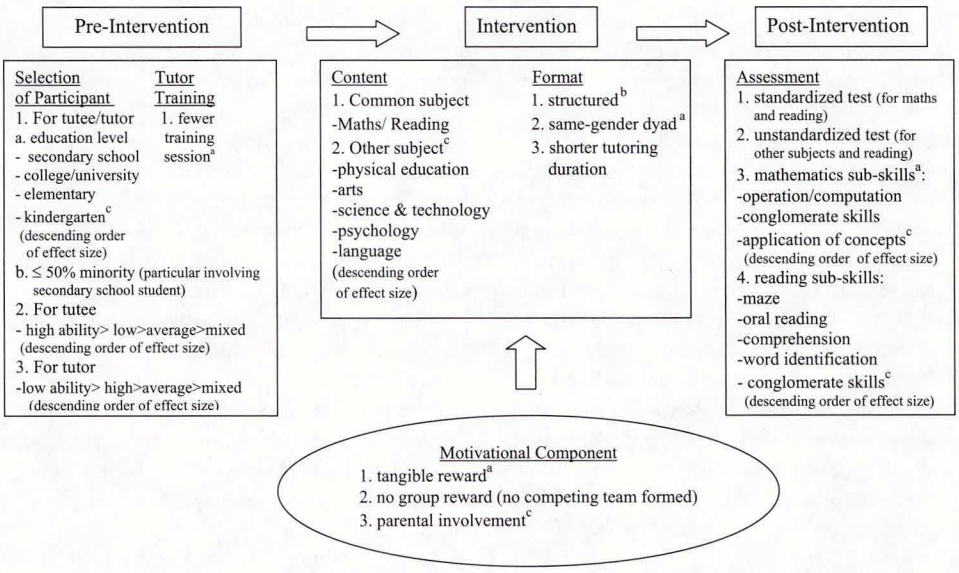

Figure 4. Preliminary empirical model of crucial determinants of best practice for peer tutoring on academic achievement. All the moderators display significant effect on the achievement in fixed effects model except those marked with superscript *b*. [a] These moderators also show significant effect under mixed effects model. [b] The moderator shows significant effect only under mixed effects model. [c] Caution should be taken when interpreting the analyses since the number of studies is very small (*n*s = 2–6).

Regarding the motivational elements, the following elements can be more effective: the adoption of tangible rewards, no formation of competing teams for group rewards, no regular formation of new competing teams, and engaging parents.

During the postintervention phase, the use of standardized or unstandardized tests for the postintervention (i.e., posttest) assessment is dependent on the subject content since there was significant interaction between subject matter and nature of test administered for controlling for author bias (using standardized test) When adopting standardized test, the use of mathematics and reading as subject content displayed larger effect size than group of other subjects. On the other hand, when adopting unstandardized test, the use of the group of other subjects and reading as subject content displayed larger effect size than mathematics. Regarding the assessment of the subskills of mathematics achievement, outcome measure on operation/computation displayed the largest effect size, followed by conglomerate skills, and application of concept. For reading achievement, outcome measure on maze (replacement of words/phrases) displayed the largest effect size, followed by oral reading, comprehension, word identification, and conglomerate skills.

From the teacher training perspective, this empirical model serves as a reference that educational trainers can use to incorporate the program features discussed to produce optimal peer tutoring effectiveness during training for preservice and in-service teachers.

Limitations and Future Research

While the methodologies adopted in the present meta-analysis provide promising directions for research assessing the impacts of peer tutoring on academic achievement, it is also plausible to adopt these techniques to explore which program features constitute the "best practice" of peer tutoring in other domains, such as motivation and nonacademic outcomes. Additionally, because the moderators identified in this meta-analysis are comprehensive, it can provide a framework for conducting further meta-analyses on peer tutoring in other domains or on all other peer-involved interventions such as peer counselling and peer-mediated interventions.

A major limitation of the present investigation is the considerable number of missing studies, which did not report sufficient data for computing the effect size. This limitation implies that future research on peer tutoring should be conducted using an appropriate methodology that can provide sufficient data for computation of effect sizes.

Another key concern is the uneven distribution of the number of studies for some subgroups because such a distribution would decrease the reliability of the results of the moderator analyses based on a small number of studies. Specifically, only two studies for kindergarten and nine studies for college or university examined for the participant's education level. For SES, only two studies came from mixed SES samples. For subject content, a small number of studies were found for subjects other than the two commonly tutored subjects (i.e., mathematics and reading). For mathematics and reading subskills, a small number of studies were used for application of mathematic concept and conglomerate reading skills. For parental involvement, only two studies involved parent participation. For dyad assignment, only two studies allowed tutees to choose their own partners. Correlation analyses between the effect size and the number of studies can be conducted to evaluate whether the imbalance in the number of studies displays an undue impact on the effect size (D'Mello, 2013). Although there is an unbalanced distribution of studies concerning

some subgroups in certain moderators, it was revealed that the effect size was not affected by the number of studies for a particular moderator. For example, for the educational level of participants, the effect size was not significantly correlated with the number of studies ($r = -.06$, $p = .94$), which indicates that the imbalance in the number of studies did not have a significant impact on the result. The insignificant correlation with number of studies is also observed for the SES, parental involvement, dyad assignment, subject content, and mathematics and reading subskills. However, additional research should be conducted in these subgroup variables to obtain more evidence and increase the statistical power of the analyses.

In addition, the findings of the present meta-analysis indicated that most of the significant moderators of academic achievement contain a heterogeneous cell, as indicated by a significant Qw, in the within-study homogeneity in the fixed effects model. This finding suggests that the effect sizes of these moderators varied across the studies in the present investigation. It is consistent with the findings of previous meta-analyses involving either peer-assisted learning (e.g., Rohrbeck et al., 2003) or other intervention programs such as counselling (e.g., Whiston, Lee, Rahardja, & Eder, 2011). However, future studies may address these issues by recombining certain subgroups based on strong theoretical ground to obtain within-study homogeneity.

References

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12,* 385–390. doi:10.1111/1467-9280.00371

Bar-Eli, N., & Raviv, A. (1982). Underachievers as tutors. *Journal of Educational Research, 75,* 139–143.

Beatty, L. S., Madden, R., Gardner, E. F., & Karlsen, B. (1986). *Stanford Diagnostic Mathematics Test* (3rd ed.). San Antonio, TX: Psychological Corporation

Bierman, K. L., & Furman, W. (1981). Effect of role and assignment rationale on attitudes formed during peer tutoring. *Journal of Educational Psychology, 73,* 33–40. doi:10.1037/0022-0663.73.1.33

Borenstein, M. (2005). Software for publication bias. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 193–220). West Sussex, England: Wiley.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-Analysis (Version 2) [Computer software]. Englewood, NJ: Biostat.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Education outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248. doi:10.3102/00028312019002237

Cook, S. B., Scruggs, T. E., Mastropieri, M. A., & Casto, G. C. (1985). Handicapped students as tutors. *Journal of Special Education, 19,* 483–492. doi:10.1177/002246698501900410

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Dion, E., Roux, C., Landry, D., Fuchs, D., Webby, J., & Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science, 12,* 70–79. doi:10.1007/s11121-010-0182-5

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105,* 1082–1099. doi:10.1037/a0032674

Durlak, J. A. (1995). Understanding meta-analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 319–352). Washington, DC: American Psychological Association.

Durlak, J. A., & Lipsey, M. W. (1991). A practitioner's guide to meta-analysis. *American Journal of Community Psychology, 19,* 291–332.

Duval, S. (2005). The trim and fill method. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127–144). West Sussex, England: Wiley.

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method for accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill. A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 124–174). Mahwah, NJ: Erlbaum.

Eccles, J. S., Jacobs, J. E., & Harold, R. E. (1990). Gender role stereotypes, expectancy effects, and parents' socialization of gender differences. *Journal of Social Issues, 46,* 183–201. doi:10.1111/j.1540-4560.1990.tb01929.x

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at-risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92,* 605–619. doi:10.1037/0022-0663.92.4.605

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education.* New York, NY: Routledge. doi:10.4324/9780203451496

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17,* 120–128. doi:10.1037/a0024445

Fogarty, J. L., & Wang, M. C. (1982). An investigation of the cross-age peer tutoring process: Some implications for instructional design and motivation. *Elementary School Journal, 82,* 450–469. doi:10.1086/461281

Fuchs, L. S., Fuchs, D., & Hamlett, C. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83,* 103–110.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28,* 617–641.

Fuchs, L. S., Fuchs, D., Powell, S. R., Seethaler, P. M., Cirino, P. T., & Fletcher, J. M. (2008). Intensive intervention for students with mathematics disabilities: Seven principles of effective practice. *Learning Disability Quarterly, 31,* 79–92.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford Achievement Test* (7th ed.). San Antonio, TX: Psychological Corporation.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1987). *Stanford 7 Plus (Primary 1).* San Antonio: Harcourt, Brace, Jovanovich.

Ginsburg-Block, M. D., & Fantuzzo, J. (1997). Reciprocal peer tutoring: An analysis of "teacher" and "student" interactions as a function of training and experience. *School Psychology Quarterly, 12,* 134–149. doi:10.1037/h0088955

Hagley, F. (2002). *Suffolk Reading Scale II: Teacher's guide.* London, England: GL Assessment.

Hammill, D. D., & Larsen, S. C. (1988). *Test of Written Language–2.* Austin, TX: Pro Ed.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128. doi:10.2307/1164588

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. doi:10.1037/0033-2909.92.2.490

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504. doi:10.1037/1082-989X.3.4.486

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539–1558. doi:10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Quantifying heterogeneity in a meta-analysis. *BMJ: British Medical Journal, 327,* 557–560. doi:10.1136/bmj.327.7414.557

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11,* 193–206. doi:10.1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22,* 255–278.

Jastak, J. F., & Jastak, S. R. (1965). *Wide Range Achievement Test.* Wilmington, DE: Guidance Associates.

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47–62. doi:10.1037/0033-2909.89.1.47

Karlsen, B., & Gardner, E. F. (1986). *Stanford Diagnostic Reading Test: Manual for interpreting—Brown level* (3rd ed). New York, NY: Psychological Corporation.

Limbrick, E., McNaughton, S., & Glynn, T. (1985). Reading gains for underachieving tutors and tutees in a cross-age tutoring programme. *Journal of Child Psychology & Psychiatry, 26,* 939–953. doi:10.1111/j.1469-7610.1985.tb00608.x

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Mackiewicz, S. M., Wood, C. L., Cooke, N. L., & Mazzotti, V. L. (2010). Effects of peer tutoring with audio prompting on vocabulary acquisition for struggling readers. *Remedial and Special Education, 32,* 345–354. doi:10.1177/0741932510362507

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin, 28,* 1183–1193. doi:10.1177/01461672022812004

Mathes, K. L., & Fuchs, L. S. (1994). The efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review, 23,* 59–80.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four vendors. *Personnel Psychology, 59,* 927–953. doi:10.1111/j.1744-6570.2006.00059.x

McMaster, K. L., Fuchs, D., & Fuchs, L. S. (2006). Research on peer-assisted learning strategies: The promise and limitations of peer-mediated instruction. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 22,* 5–25. doi:10.1080/10573560500203491

Neale, M. D. (1997). *Neale Analysis of Reading Ability–Revised.* Windsor, United Kingdom: NFER-Nelson.

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41,* 181–206. doi:10.1207/s15326985ep4103_4

Orwin, R. G. (1983). A Fail-Safe N for effect size. *Journal of Educational Statistics, 8,* 157–159. doi:10.2307/1164923

Overton, R. C. (1998). A comparison of fixed effects and mixed (random effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379. doi:10.1037/1082-989X.3.3.354

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin, 134,* 270–300. doi:10.1037/0033-2909.134.2.270

.Pereira-Laird, J. A., & Deane, F. P. (1997). Development and validation of a self-report measure of reading strategy use. *Reading Psychology, 18,* 185–235.

Pham, B., Platt, R., McAuley, L., Klassen, T. P., & Moher, D. (2001). Is there a "best" way to detect and minimize publication bias? An empirical evaluation. *Evaluation and the Health Profession, 24,* 109–125.

Polirstok, S. R., & Greer, R. D. (1986). A replication of collateral effects and a component analysis of a successful tutoring package for inner-city adolescents. *Education & Treatment of Children, 9,* 101–121.

Prescott, G. A., Balow, I. H., Hogan, T. P. & Farr, R. C. (1978). *Metropolitan Achievement Tests: Teacher's manual for administering and interpreting.* New York, NY: Psychological Corporation.

Raudenbush, S. W. (1994). Random effects model. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York, NY: Russell Sage Foundation.

Reid, D. K., Hresko, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability* (3rd ed.). Austin, TX: Pro-Ed.

Roach, J. C., Paolucci-Whitcomb, P., Meyers, H. W., & Duncan, D. A. (1983). The comparative effects of peer tutoring in math by and for secondary special needs students. *Pointer, 27,* 20–24.

Robinson, D. R., Schofield, J. W., & Steers-Wentzell, K. L. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review, 17,* 327–362. doi:10.1007/s10648-005-8137-2

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology, 95,* 240–257. doi:10.1037/0022-0663.95.2.240

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi:10.1037/0033-2909.86.3.638

Rosenthal, R. (1984). *Meta-analytic procedures for social science research.* Beverley Hills, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500–504. doi:10.1037/0033-2909.92.2.500

Sáenz, L. M., Fuchs, L. S., & Fuchs, D. (2005). Peer-assisted learning strategies for English language learners with learning disabilities. *Exceptional Children, 71,* 231–247. doi:10.1177/001440290507100302

Scruggs, T. E., & Ritcher, L. (1988). Tutoring learning disabled students: A critical review. *Learning Disability Quarterly, 11,* 274–286. doi:10.2307/1510772

Simmons, D. C., Fuchs, D., Fuchs, L. S., Hodge, J. P., & Mathes, P. G. (1994). Importance of instructional complexity and role reciprocity to classwide peer tutoring. *Learning Disabilities Research & Practice, 9,* 203–212.

Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice.* Englewood Cliffs, NJ: Prentice Hall.

Soeken, K. L., & Sripusanapan, A. (2003). Assessing publication bias in meta-analysis. *Nursing Research, 52,* 57–60. doi:10.1097/00006199-200301000-00009

Spencer, V. G., & Balboni, G. (2003). Can students with mental retardation teach their peers? *Education and Training in Developmental Disabilities, 38,* 32–61.

Sprinthall, N. A., & Scott, J. R. (1989). Promoting psychological development, math achievement, and success attribution of female students through deliberate psychological education. *Journal of Counseling Psychology, 36,* 440–446. doi:10.1037/0022-0167.36.4.440

Stecker, P. M., Fuchs, L. S., & Hamlett, C. (1992). *Technical features of mathematics operations and applications progress monitoring measures*. Unpublished manuscript, Peabody College, Vanderbilt University, Nashville, TN.

Sutton, A. (2005). Evidence concerning the consequences of publication and related bias. In H. Rothstein, A. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 175–192). West Sussex, England: Wiley.

Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology, 66,* 21–36. doi:10.1037/0022-3514.66.1.21

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review, 31,* 498–513.

Topping, K. (1987). Peer tutored paired reading: Outcome data from ten projects. *Educational Psychology, 7,* 133–145. doi:10.1080/0144341870070206

Topping, K. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education, 32,* 321–345. doi:10.1007/BF00138870

Topping, K. (2005). Trends in peer learning. *Educational Psychology, 25,* 631–645. doi:10.1080/01443410500345172

Topping, K. J., Miller, D., Murray, P., Henderson, S., Fortuna, C., & Conlin, N. (2011). Outcomes in a randomised controlled trial of mathematics tutoring. *Educational Research, 53,* 51–63. doi:10.1080/00131881.2011.552239

Topping, K. J., & Whiteley, M. (1993). Sex differences in the effectiveness of peer tutoring. *School Psychology International, 14,* 57–67. doi:10.1177/0143034393141004

Topping, K. J., Thurston, A., McGavock, K., & Conlin, N. (2012). Outcomes and processes in reading tutoring. *Educational Research, 54,* 239–258. doi:10.1080/00131881.2012.710086

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Underwood, J., Underwood, G., & Wood, D. (2000). When does gender matter? Interactions during computer-based problem-solving. *Learning and Instruction, 10,* 447–462. doi:10.1016/S0959-4752(00)00008-6

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relations between self-beliefs and academic achievement: A systematic review. *Educational Psychologist, 39,* 111–133. doi:10.1207/s15326985ep3902_3

Vernon, P. E. (1969). *The standardization of a graded word reading test*. London, England: University of London Press.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

Wentzel, K. R. (1999). Socio-motivational processes and interpersonal relationships: Implications for understanding motivation at school. *Journal of Educational Psychology, 91,* 76–97. doi:10.1037/0022-0663.91.1.76

Whiston, S. C., Lee, T. W., Rahardja, D., & Eder, K. (2011). School counseling outcome: A meta-analytic examination of interventions. *Journal of Counseling & Development, 89,* 37–55.

Woodcock, R. N. (1973). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Received October 23, 2013
Revision received June 29, 2014
Accepted June 30, 2014 ■