LEU05421

Are Peer Tutoring Programs Effective in Promoting Academic Achievement and Self-Concept in Educational Settings: A Meta-Analytical Review

Charles K. C. Leung, Herbert W. Marsh, and Rhonda G. Craven

SELF Research Centre, University of Western Sydney, Australia

Abstract

Previous meta-analyses of the effects of peer tutoring on academic achievement and self-concept have been confined to specific populations particular subject content, plagued with methodological flaws in meta-analysis or did not capture theoretical and methodological advances in self-concept research. An updated, comprehensive meta-analysis evaluating the effect of peer tutoring on academic achievement and self-concept was conducted, based on current advances in meta-analysis methodology, engaging a wide range of participants, subject content and adopting a construct validity approach to the study of self-concept intervention effects. The findings demonstrated that peer-tutoring programs impacted positively on academic achievement (unweighted mean ES =0.81, SD =0.79; weighted ES =0.65, $p<.05$, 95% confidence interval = 0.59 - 0.71) and self-concept (unweighted mean ES =0.82, SD =0.80; weighted ES =0.88, $p< .05$, 95% confidence interval =0.69 - 1.07). On the basis of the construct validity approach, homogeneity analyses ($Q_B=57.00$, $k=1$, $p<.05$) showed that peer tutoring programs had greater effect on target domains of self-concept consistent with the design of the intervention than on non-target self-concept. The implications of these findings are discussed.

Peer tutoring has been commonly implemented in education settings. Research has shown that peer tutoring has a positive impact on academic outcomes such as reading (e.g. Klingner & Vaughn, 1996), mathematics (e.g. Fuchs, Fuchs, & Karns, 2001), spelling and other subjects (e.g. Riggio, Fantuzzo, Connelly & Dimeff, 1991). Several reviews have employed meta-analysis to systemically review the effect of peer tutoring empirically. However, many of these meta-analyses are dated. For example, the meta-analytic review done by Cohen, Kulik and Kulik (1982) was confined to literature published prior to 1980. Moreover, it did not capture the methodological advances in meta-analysis including: calculating effect size employing correcting procedures due to sample size, utilizing weighting procedures that account for sample size, selecting the appropriate unit of analysis, dealing with outliers by windsorizing effect size, and using homogeneity analyses to examine moderator variable and group differences. Of the recent meta-analyses examining peer tutoring which adopt at least some of the current advances in methodology in meta-analysis, these are confined to certain populations such as elementary school children (e.g. Rohrbeck, Ginsburg-Block, Fantuzzo & Miller, 2003) or are based upon using adult (teachers, adult volunteers or college students) as tutors instead of using peers (Elbaum, Vaughn, Hughes & Moody, 2000). The aim of this study is to synthesize previous research pertaining to peer tutoring and address limitations of previous meta-analytic research by: including studies that examine a range of subject content and a wide range of participants, and adopting current methodological advances in meta-analysis in order to critically evaluate the impact of academically–orientated peer tutoring programs on academic outcomes. Since early meta-analyses have

suggested that that academically-orientated peer tutoring programmes impact positively on academic achievement, it was hypothesized that these results would be replicated in this updated meta-analysis. Following the common practice of meta-analysis, homogeneity tests were also conducted to assess whether the effect of peer tutoring is moderated by certain mediating variables.

A positive self-concept is considered as a desirable outcome in educational settings. Hence, it is useful to test the effect of peer tutoring on self-concept. The meta-analysis conducted by Cohen, Kulik and Kulik (1982) showed that tutoring had little or no effect on self-esteem. Cook, Scruggs, Mastropieri and Casto (1985) also concluded same result. However, as mentioned previously, these studies did not employ recent methodological advances in meta-analysis. Importantly, these studies also did not account for the multidimensionality of self-concept when examining the effect of peer tutoring on self-concept.

Shavelson, Hubner and Stanton (1976) posited that self-concept is multifaceted instead of unidimensional in nature. Described in the Shavelson model is a general self-concept defined by academic and non-academic self-concepts. Academic self-concept is further divided into self-concepts in particular content areas whereas non-academic self-concept is divided into social, physical, and emotional self-concepts. The multidimensionality of self-concept has been supported by numerous factor analytic studies (e.g., Harter, 1982; Marsh, Barnes, & Hocevar, 1985; Marsh, Parker, & Barnes, 1985) and construct validity reviews (e.g., Byrne, 1984; Marsh & Shavelson, 1985). Hence, recent self-concept research emphasises the multidimensionality and domain-specificity of self-concept. Advances in self-concept theory and research demonstrate that self-concept cannot be adequately understood if its dimensionality is ignored (Marsh & Craven, 1997; in press). However, previous meta-analyses examining the impact of academically-orientated peer tutoring programmes have focused on a unidimensional self-concept construct and as such have not accounted for the multidimensionality of the construct. This has resulted in paradoxical findings. Craven, Marsh and Burnett (2003) advocate a construct validity approach to the study of self-concept intervention effects and have demonstrated that facets of self-concept logically targeted by self-concept enhancement interventions have been enhanced. Hence, in the present investigation, it is anticipated that the construct validity approach to the study of intervention effects will be supported whereby target facets of self-concept most relevant to the goals of the interventions will display a greater effect size compared to non-target facets of self-concept less relevant to the goals of the interventions.

Method

Research Design

The present investigation was comprised of a meta-analysis designed to test the effects of academically-orientated peer tutoring programmes on academic achievement and self-concept. Academically-orientated peer tutoring programmes were characterized as including “a system of instruction in which learners help each other and learn by teaching” (Goodlad & Hirst, 1989, p. 13) or “a more able child

helping a less able child in a cooperative working pair carefully organized by a teacher" (Topping, 1989, p. 489). Regarding the self-concept, the research design was based upon a construct validity approach to the study of intervention effects (see Craven et al., 2003) whereby the impact of academically-orientated interventions was evaluated in relation to both target facets of self-concept most relevant to the goals of each intervention and non-target facets that were less relevant to the goals of the intervention.

Procedures

The meta-analysis followed the guidelines established in previous meta-analysis research (Glass, McGraw, & Smith, 1981; Hedges & Olkin, 1985). It involved the following procedures: Literature search, coding of studies and computation of effect sizes.

Selection criteria and procedures. For the literature search, key terms including peer tutoring and peer tutor were used to search the following online databases: PsycLit, Educational Resources Information Centre (ERIC) and Dissertation Abstracts. The following criteria were set for the eligibility of studies: (1) The study was peer-reviewed and published in 2003 or before; (2) The form of peer tutoring needed to take place in school setting; (3) Participants were kindergarten, primary, high school, college or university students; (4) The targeted subject matter was academic; (5) Outcome data available in the article needed to be amenable to the computation of effect sizes; and (6) Studies were selected based upon experimental designs that included only an experimental group with pretest and posttest scores or both a control and experimental group.

Coding of studies. The code sheet utilized in a previous meta-analysis on tutoring (Cohen et al., 1982; Elbaum, Vaughn, Hughes & Moody, 2000) was used initially as a basis to develop a coding sheet. Specifically this coding schema included: (1) report information; (2) characteristics of participants; (3) methodology; (4) treatment features; and (5) outcomes assessment. Coding of all eligible studies was done by the first author whereas half of randomly selected studies were coded by a second coder who is a secondary school teacher with a master's degree in education. The code sheets completed by the two coders were compared to confirm the inter-rater reliability. The inter-rater reliability was calculated as the percentage agreement between codes assigned by two coders.

Data analysis. ***Unit of analysis.*** Since independent samples were the primary unit of analysis each study contributed one independent sample to the analysis and one effect size was calculated. If a study reported findings separately for different subgroups such as a low-achiever, average-achiever and high-achiever, these effect sizes were calculated separately, however, only one overall effect size was calculated for that study by averaging these effect sizes. Similarly, if a study reported findings separately for different subsets of one type of achievement measure (e.g. oral and comprehension of reading), it would contribute two effect sizes for that study. Again, only one overall effect size was calculated for similar achievement domains by averaging these effect sizes.

Computation of standardized mean difference effect size. Standardized effect size was calculated by dividing the difference between the treatment and control group means with the pooled standard deviation of the two groups (Hedges, 1981). For those studies in which only test statistics such as *t* or *F* was available, different formulas were used for computation of effect size (Lipsey & Wilson, 2000). For those studies only reporting gain score (pretest, posttest or standard deviation), standardized gain score effect size (Hedges, 1981) was calculated. For computing the effect size for pretest-posttest designs in which pretest and posttest data for experimental group is given without control, the formula suggested by Becker (1988) was used (see Appendix 1). Since effect size is positively biased with small samples (Hedges & Olkin, 1985), the standard method of correcting such bias was conducted (see Appendix 1).

Computation of mean effect sizes. Since magnitude of effect size varies with the sample size, it is inappropriate to treat each effect size as an equal estimator of the underlying population mean effect size. Therefore, it is essential to weight each effect size by its sample size by following the standard meta-analytic practice (Hedges & Olkin, 1985). A mean weighted effect size was calculated based on the principle that the greater sample size of the study, the greater weight was given to effect size of that study (Lipsey & Wilson, 1996). Mean weighted effect size was calculated by multiplying each effect size by its weight in which was computed as the inverse of the variance of the effect size estimate (Cooper, 1989). Hence, the mean effect size is calculated by dividing the summation of all weighted effect sizes by the summation of the weights associated with each effect size.

Homogeneity analysis. To examine whether each set of effect sizes are each estimating the same population effect size, a homogeneity test was conducted (Hedges, 1982; Rosenthal & Rubin, 1982). Homogeneity tests help to evaluate whether the variability of the effect sizes within a particular category was accounted for by sampling error alone or special coded feature of the effect size of the particular set. In a homogeneity distribution, it was expected that the dispersion of effect sizes within a particular category around mean effect size is no greater than would be expected by sampling error alone. Conversely, dispersion of effect sizes within a particular category around mean effect size is greater than would be expected by chance if the homogeneity test is rejected. Hence, the dispersion of effect size within a particular set does not estimate the same population mean (Lipsey & Wilson, 1996). Homogeneity tests were examined by using the Q statistic in which the distribution is similar to chi-square with k-1 degrees of freedom where k is the number of effect sizes (Hedges & Olkin, 1985). Categorical approach was adopted (Hedges, 1982) in the present study to determine the relation between the special coded feature of the study and the magnitude of the effect sizes. Between-group homogeneity statistic, Q_B and within-group homogeneity statistic, Q_W (Hedges & Olkin, 1985) were calculated. A significant Q_B indicates that the average effect size differs over groups and the group variable is a significant moderator of outcome whereas a non-significant Q_w suggests that the variable appropriately groups studies into homogenous subcategories and the effect sizes under the variable are consistent across the studies.

Study Characteristics

Search Outcome

A total of 76 articles were identified that met for the criteria for inclusion. Of these, one article used mixed strategies whereas 8 articles did not report sufficient data amenable to the computation of effect size. These articles were excluded from the present review. Hence, 68 articles were retained for further analysis.

Coding Reliability

Two coders used the coding sheet to code the articles. In order to have a common understanding of the items in coding sheet, two pilot coding sessions were held to discuss any disparity on the coding sheet. After concluding consensus, each coder did the coding separately. Whilst the first author coded all 68 studies, the second coder coded a sample of 34 randomly selected studies. Inter-rater reliability was calculated by using percentage agreement and kappa coefficient, where appropriate for these studies. Average percentage agreement and kappa coefficient for variables were 92% and 0.94 respectively.

Profile of Effect Size

There were 68 studies yielding a total of 201 effect sizes, with 167 effect sizes for achievement and 34 effect sizes for self-concept. Examining the distribution of the unweighted effect size to reveal any outliers was important. Outliers were defined as those effect sizes that were more than three interquartile ranges beyond the 75th percentile (called positive outliers) or more than three interquartile ranges below the 25th percentile (called negative outliers) based on the Tukey's definition (Tukey, 1977). These outliers were winsorized by setting their values to three interquartile ranges beyond the 75th percentile for positive outliers and 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 size was still accounted for (Tabachnick & Fidell, 2001; Durlak & Lipsey, 1991).

For achievement in the present investigation, 6 positive effect sizes from 3 studies met this criterion and were considered as outliers. These effect sizes were winsorized by setting their values to three interquartile ranges beyond the 75th percentile. After winsorizing these 6 extreme effect sizes, the unweighted mean effect size for the 167 effect size for academic achievement was 0.83 (SD=1.30). Using the 68 independent samples as the unit of analysis, the unweighted mean effect size was 0.81 (SD=0.79). Regarding self-concept, there were no outliers and the unweighted mean effect size for the 34 effect size was 0.58 (SD=0.74). Using the 8 independent samples as the unit of analysis, the unweighted mean effect size was 0.82 (SD=0.80).

Examining the distribution of the sample size for checking outliers was necessary since weighting of effect size was based on sample size and hence, extreme large sample sizes would have undue effect on the findings. A total of 3 studies for achievement and 2 studies for self-concept met the criteria as outliers. These samples were winsorized by setting their values to three interquartile ranges beyond the 75th percentile. After winsorizing these extreme sample sizes, the average sample size for the present investigation for academic achievement was 55.93 (SD=49.39). Regarding the self-concept, the average sample size was 36.24 (SD=33.20).

As mentioned previously, each effect size was weighted by the inverse of their variance (Cooper & Hedges, 1994; Hedges & Olkin, 1985) since studies with large sample size provide more reliable estimation of population effect size. The weighted mean effect size for independent samples in the present investigation for achievement was 0.65. Regarding the self-concept, the weighted mean effect size was 0.88.

Publication Bias

Since only published articles were included in the meta-analysis, it was likely that the effect size was overestimated because published material would report only significant findings or findings in support of the hypothesis. To estimate whether the bias occurred, a fail-safe sample size (FSN) was calculated (Rosenthal, 1979). A fail-safe sample size (FSN) is the number of unpublished studies with no significant effect size that would be needed to overturn the overall effect size to an unimportant level (Orwin, 1983). The FSN is:

$K_o = k [ES_k/Es_c -1]$

K_o is the number of studies with a value of zero need to reduce the mean effect size to Es_c
k is the number of studies in the present investigation
Es_c is the criterion effect size level
ES_k is the weighted mean effect size

In the present study, the overall weighted mean effect size for academic achievement was 0.65. It was defined that a trivial effect size was 0.10. Hence, the FSN was equal to 374. In other words, an additional 374 studies with no effects would be needed to decrease the overall effect size of the present study to an insignificant level. Therefore, it was rather unlikely to expect such large number of unpublished studies with null results. Moreover, using published material has the advantage that the methodology employed is presumably more rigorous.

Regarding the self-concept, the overall weighted mean effect size was 0.88. It was defined that a trivial effect size was 0.10. Hence, the FSN was equal to 62. It means that 62 additional studies with no effects would be required. Since there were only 8 out of 68 studies (12%) found for self-concept in the present investigation, it was deduced that about 500 articles to be identified for getting 62 studies for self-concept would be required. Hence, it was rather unlikely that there were such large number of unpublished studies with null results.

Demographic Characteristics of Students

Most of the studies (85.3%, *n*=55) came from 1980-1999. Most of the studies (95.6%, *n*=65) reported the grade level of the tutees ranging from kindergarten to university and only 3 (4.4%) studies did not report the grade level of tutees. A total of 72.1% of the studies (*n*=49) engaged tutees in elementary grade, 13.2% (*n*=9) of the studies engaged tutees in middle school grade, and only a few studies engaged tutees in kindergarten grade (2.9%, *n*=2), upper secondary school grade (2.9%, *n*=2), and in college or university grade (4.4%, *n*=3). A similar pattern was found for the tutors. Most of the studies (97.1%, *n*=66) reported the grade level of the tutors ranging from kindergarten to university and only 2 (2.9%) studies did not report the grade level of tutors. A total of 66.2% (*n*=45) of the studies engaged tutors in elementary grade, 16.2% (*n*=11) of the studies engaged tutors in middle school grade and only a few studies engaged tutees in kindergarten grade (2.9%, *n*=2), 5.9% (*n*=4) in upper secondary school grade, and 4.4% (*n*=3) in college or university grade. In addition, one study engaged tutors in both elementary and college grade.

Most of the studies reported the age of tutee (75%, *n*=51) and tutors (76.5%, *n*=52), with tutees and tutors ranging from 5.9 to 35.5 years. Most of the studies engaged tutees (58.8%, *n*=40) and tutors (50.0%, *n*=34) aged 5 to12 years. Since most of the studies (70.6%, *n*=48) adopted same-age peer tutoring, the mean age of tutor and tutee was similar. The mean age of tutee was 11.57 (SD = 6.16) and tutor was 12.07 (SD = 5.93).

A total of 41.2% (*n*=28) of the studies reported the socio-economic status (SES) of participants while 58.8% (*n*=40) did not report this. Also 22.1% (*n*=15) of the studies included participants of low SES, 19.1% (*n*=13) of the studies with mixed SES, and none of the studies reported participants having middle or high SES. Regarding the ethnicity of participants, half of the studies reported the ethnicity of participants while half did not report the data. Some 5.9% (*n*=4) of the studies included Caucasian participants, 8.8% (*n*=6) of the studies with Afro-American participants, 5.9% (*n*=4) of the studies with Black participants, 1.5% (*n*=1) of the studies with Asian participants, 20.6% (*n*=14) of the studies with participants of mixed ethnicity and 7.4% (*n*=5) of the studies with participants of other ethnicity.

A total of 75% (*n*=51) of the studies provided reported the academic ability of the tutees and 25 % (*n*=17) did not. Also 22.1 % (*n*=15) of the studies engaged tutees of low ability, 23.5% (*n*=16) tutees with special need, 29.4% (*n*=20) tutees with mixed ability and none of the studies characterized tutees as having average ability or high ability. For tutors, a similar trend was revealed since most of the studies adopted same-age peer tutoring. Such that 72.1% (*n*=49) of the studies provided reported the academic ability of the tutors and 27.9 % (*n*=19) did not provide this data. A total of 20.6% (*n*=14) of the studies included tutors of low ability, 20.6% (*n*=14) tutors with special need and 27.9% (*n*=19) tutors with mixed ability. Only 2.9% (*n*=2) of the studies included tutors with high ability and none of studies included tutors with average ability.

The mean number of tutees was 57.85 (SD= 60.49) ranging from 4 to 282 whereas the mean number of tutors was 56.94 (SD= 61.00) ranging from 3 to 282. A total of 41.2 % of the studies reported the gender of the participants.

Of the target sample, 60.3% (*n*=41) of the studies selected participants on an individual basis and 38.2% (*n*=26) on a whole-class basis. Only 1 study selected the participants on a whole-class sample of convenience basis and no study reported that the participants were self-selected.

Characteristic of Methodology Parameters

As mentioned above, 48.5% of the studies (*n*=33) were same-age reciprocal peer tutoring and 22.1% of the studies (*n*=15) conducted same-age non-reciprocal peer tutoring, 25.0% (*n*=17) adopted cross-age peer tutoring and only 3 studies (4.5%) implemented mixed mode of peer tutoring. Most of the studies (85.3%, *n*=58) were monitored or led by a teacher and only a few studies (14.7%, *n*=10) were monitored by other people. While most of the studies (97.1%, *n*=66) did not involve parents, one study engaged parents in the intervention. A total of 76.5% (*n*=52) of the studies adopted structural tutoring and only a few studies (19.1%, *n*=13) did not. A fidelity check of the intervention was undertaken for 69.1% (*n*=47) of the studies. A total of 63.2% (*n*=43) of the studies controlled for author bias by using standardized tests, 33.8% (*n*=23) of studies did not, and 2 studies adopted both standardized and unstandardized tests. In addition, most of the studies (95.6%, *n*=65) controlled for instructor bias whereas only 2.9% (*n*=2) of studies did not. Regarding tutor training, most of the studies (85.3%, *n*=58) reported that there was training and 14.7% (*n*=10) did not conduct training. Also 23.5% (*n*=16) of the studies used tutoring as substitute to the original classroom instruction while most of the studies (76.5%, *n*=52) did not. Furthermore, most of the studies (82.5%, *n*=56) used tutoring as supplement to the original classroom interaction while 17.6% (*n*=12) did not.

Most of the studies (73.5%, *n*=50) adopted either a control group or comparison group and 26.5% (*n*=18) of the studies did not. For those with comparison or control group, 5.9% (*n*=4) of the studies adopted equivalent comparison group, 17.6% (*n*=12) used a non-equivalent comparison group, 23.5% (*n*=16) assigned individuals to control and experimental groups, 23.5% (*n*=16) assigned participants to control and experimental groups based upon a group or class basis and 3% (*n*=2) adopted two types of assignment of participants whereas none of the studies used match-subject design. Most of the studies (95.6%, *n*=65) administered pretest and posttest measures whereas only few studies (4.4%, *n*=3) also included a follow-up measure.

Characteristic of Intervention Parameters

As described previously, most of the studies (85.3%, *n*=58) reported that there was training for the tutor. Of these 47.1% (*n*=32) of the studies reported the duration of tutor training with an average of 3.78 sessions and 38.2% (*n*=26) did not. Also 41.2% (*n*=28) of the studies reported the length of each session of tutor training with an average of 62.38 minutes per session and 44.1% (*n*=30) did not.

For the intervention, 69.1% (*n*=47) of the studies reported the number of sessions of tutoring with an average of 2.98 sessions per week and 30.9% (*n*=21) did not. Also 73.5% (*n*=50) of the studies reported the length of each session of tutoring with an average of 30.05 minutes per session and 26.5% (*n*=18) did not. A

total of 80.9% (*n*=55) of the studies reported duration of tutoring with an average of 16.53 weeks and 19.1% (*n*=13) did not. For intervention setting, most of the studies (91.2%, *n*=62) reported the data and a few studies (8.8%, *n*=6) did not. In addition, 66.2% (*n*=45) reported that the intervention took place during school lesson, 23.5% (*n*=16) took place in other period whereas only 1 study took place after school (1.5%).

Most of the studies (36.8%, *n*=25) conducted an intervention targeting reading, followed by mathematics (25%, *n*=17) and other subjects (27.9%, *n*=19). There were 2 studies conducted that targeted intervention on both on mathematics and reading, and 1 study targeted mathematics and another subject and 2 studies targeted mathematics, reading and other subjects. For the instrument for measuring achievement outcomes, a larger proportion of the studies for mathematics (25 %, *n*=17) and reading (39.7 %, *n*=26) used a standardized measure whereas only small number of studies (mathematics: 7.4 %, *n*=5; reading 2.9 %, *n*=2) utilized a researcher-devised measure for the study, and only 1 study used both these types of measures for reading. For measuring achievement in other subjects, a larger number of studies (20.6 %, *n*=14) utilised a researcher-devised measure and only 11.8 %, (*n*=8) used standardized measures.

Regarding the data for calculating effect size, while most of the studies (79.4%, *n*=54) were concerned with tutee, only 1 study (1.5%) examined the effects of the study on the tutor, and 13 studies (19.1%) considered both the tutee and tutor.

Results: Effect Size Analyses

Academic Achievement

Using the 68 independent samples as the unit of analysis, the unweighted mean effect size was 0.81 (SD=0.79). For overall academic achievement, the weighted mean effect size was 0.65 (*n*=66, $p<.05$, 95% CI =0.59 – 0.71). Since the confidence interval did not include zero, the effect size was significantly different from zero at the 5 percent level of significance. The positive effect size denoted that there was a greater improvement for treatment groups in comparison to control groups. Regarding the magnitude of effect size, on the basis of Lipsey and Wilson's classification (1993), it was medium to high in magnitude. This suggested that peer tutoring had a positive moderate to high effect on academic achievement.

Possible Moderators of Academic Achievement

Homogeneity analyses were conducted to assess the moderators regarding the participant parameters, methodology parameters, intervention parameters and intervention outcome parameters. Regarding the participant parameters, education level of tutee and tutor, age of tutee and tutor, ethnicity of participants, academic ability of tutee and tutor, and mode of selection of sample were all significant moderators of achievement outcome. Studies with tutees or tutors from secondary school ranging from 13 to 18 years, Caucasian students, tutees or tutors of low academic ability, and selection of participants based on an

individual basis displayed larger effect sizes in comparison to other studies (see Appendix 2).

For the methodology parameters, type of peer tutoring, structure of tutoring, nature of test administered, training to tutor, substitution to classroom instruction and nature of class interaction were all significant moderators of achievement outcome, such that same-age peer tutoring, adoption of structural tutoring, use of unstandardized test, provision of tutor training, use of tutoring as supplementary classroom instruction and to supplement to class interaction displayed larger effect sizes in comparison to other studies.

For the intervention parameters, duration of tutor training, number and length of tutoring sessions were all significant moderators of achievement outcome, such that tutor training less than or equal to 3 sessions, duration of intervention greater than or equal to 3 sessions and length of each tutoring session less than or equal to 30 minutes displayed larger effect sizes in comparison to other studies.

For the intervention outcome parameters, subject content, quality of reading and other subject measure, and data type for calculating effect size were all significant moderators of achievement outcome, such that other subject, use of measure created for the study for reading and other subject, and use of gain score and pretest-posttest change score displayed larger effect sizes in comparison to other studies.

Although Q_b was significant for these moderators, however, Q_w was also significant for most of these moderators. These results indicate that there was a greater variability of effect sizes on these moderators in each subcategory.

Self-concept

For the self-concept, similar to academic achievement, all studies (n=8) did not adopt a construct validity approach in studying the effect of peer tutoring. A total of 75% (n=6) measured only target facets of self-concept and 25% (n=2) measured only non-target facets of self-concept. Regarding the self-concept measurement instrument, 50% (n=4) of studies used a unidimensional scale including the Rosenberg scale (n=1), Piers-Harris Children scale (n=1) and other scales (n=2); 50% (n=4) of studies used Harter's Self-Perception Profile for Children multidimensional scale (n=4). All these scales were standardized measures except those created for a particular study.

Using the 8 independent samples as the unit of analysis, the unweighted mean effect size was 0.82 (SD=0.80) whereas the weighted mean effect size was 0.88 ($p<.05$, 95% CI =0.69 – 1.07). Since the confidence interval did not include zero, the effect size was significantly different from zero at the 5 percent level of significance. The positive effect size denotes that there was a greater improvement of self-concept for treatment groups over control groups. Regarding the magnitude of the effect size, on the basis of Lipsey and Wilson's classification (1993), it was high. This suggested that peer tutoring had a positive large effect on self-concept.

For the intervention's goal on self-concept domain, the weighted mean effect size for target self-concept was 1.09 (n=14, $p<.05$, 95% CI =0.93 – 1.25) and 0.18 (n=20, $p<.05$, 95% CI =0.01 – 0.35) for non-target self-concept (see Appendix 3). Homogeneity analyses showed that the focus of the intervention was a significant moderator of effect size (Q_B=57.00, k=1, $p<.05$). Hence, it suggested the effect size for

target and non-target self-concept was significantly different. Moreover, since Q_w (97.22, k=1, p <.05) was significant for target self-concept, these results indicate that the effect sizes on the target self-concept domain varied across studies.

Discussion

This updated meta-analysis using current methodology in meta-analysis replicated the results of previous meta-analyses in that peer-tutoring programs impacted positively on academic achievement regardless of the subject content and range of participants. However, there were some moderators mediating these effects which implies that researchers need to take these parameters into consideration when trying to enhance the effectiveness of peer tutoring programmes. Regarding the effect on self-concept, unlike the findings from some previous meta-analytic reviews, the results revealed that there was a high impact of peer tutoring on self-concept. Perhaps these results are due to the fact that the present study captured methodological advances in meta-analysis. Moreover, there was a greater effect of peer tutoring on target self-concept than non-target self-concept. However, caution should be taken when interpreting the data since the sample size was too small for each category. These results need further investigation using a larger sample size. Nevertheless, the present study has provided preliminary support for the positive impact of peer tutoring on self-concept. The results also have implications for designing self-concept enhancement programmes in that the multidimensionality of self-concepts as postulated in the Shavelson model needs to be considered.

About the Authors

Charles K. C. Leung is a PhD Candidate, in the SELF Research Centre. His specializations are peer support and self-concept.

Professor Herb Marsh is Director and Research Professor of the SELF Research Centre, University of Western Sydney. Largely due to his established research program into self-concept, motivation, identity and related constructs that are the focus of the SELF Research Centre, he is arguably Australia's leading researcher in both the broad disciplines of Education and Psychology

Associate Professor Rhonda Craven is Deputy Director of the SELF Research Centre – ranked 7th in the world in educational psychology, is Associate-Professor in the School of Education and Early Childhood Studies, University of Western Sydney. As an Educational Psychologist her research focuses on large-scale quantitative research studies in educational settings.
Email: r.craven@uws.edu.au

References

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257-278.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54*, 427-456.

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.

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

Cooper, H. M. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Newbury Park: CA: Sage.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Craven, R. G., Marsh, H. W. & Burnett, P. C. (2003). Cracking the self-concept enhancement conundrum: A call and blueprint for the next generation of self concept enhancement research. *International Advances in Self Research* (Vol. 1, 67-90). Greenwich, Connecticut: Information Age.

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

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.

Fuchs, L. S., Fuchs, D., & Karns, K. (2001). Enhancing kindergartners' mathematical development: Effects of peer-assisted learning strategies. *The Elementary School Journal*, 101, 495-510.

Glass, G. V., McGaw, B., & Smith. M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goodlad, S., & Hirst, B. (1989). *Peer tutoring: A Guide to Learning by Teaching*. London: Kogan Page.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development, 53,* 87-97.

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

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

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

Klingner, J. K., & Vaughn, S. (1996). Reciprocal teaching of reading comprehension strategies for students with learning disabilities who use English as a second language. *The Elementary School Journal*, 96, 275-293.

Lipsey, M.W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209.

Lipsey, M.W., & Wilson, D. B. (1996). *Toolkit for Practical meta-analysis*. Evaluation Centre: Vanderbilt University.

Lipsey, M.W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Marsh, H. W., & Craven, R. G. (1997).Academic self-concept: Beyond the dustbowl. In G. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment.* San Diego, CA: Academic Press.

Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*, 107-125.

Marsh, H. W., Barnes, J., & Hocevar, D. (1985). Self-other agreement on multidimensional self-concept ratings: Factor analysis & multitrait-multimethod analysis. *Journal of Personality and Social Psychology, 49*, 1360-1377.

Marsh, H. W., Parker, J., & Barnes, J. (1985). Multidimensional adolescent self-concepts: Their relationship to age, sex and academic measures. *American Educational Research Journal, 22*, 422-444.

Orwin, R. G. (1983). A fail-safe N for effect size. *Journal of Educational Statistics, 8,* 157-159.

Riggio, R. E., Fantuzzo, J. W., Connelly, S., & Dimeff, L. A. (1991). Reciprocal peer tutoring: A classroom strategy for promoting academic and social integration in undergraduate students. *Journal of Social Behavior and Personality*, 6, 387-396.

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.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rosenthal, R., & Rubin, D.B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500-504.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Validation of construct interpretations. *Review of Educational Research, 46,* 407-441.

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

Topping, K. (1989). Peer tutoring and paired reading: combining two powerful techniques. *The Reading Teacher, 42,* 488-494.

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

Appendix 1

(1) Computation of standardized mean difference effect size
(i)using t-value

$$ES_{sm} = t\,(n_1+n_2/n_1n_2)^{1/2}$$

ES_{sm} is the standardized mean effect size
t is the independent t-test value
n_1 and n_2 are the sample size of treatment and control group respectively

(ii)using F ratio

$$ES_{sm} = (F(n_1+n_2/n_1n_2))^{1/2}$$

ES_{sm} is the standardized mean effect size
F is the one-way ANOVA F-ratio
n_1 and n_2 are the sample size of treatment and control group respectively

(2) Computation effect size involving gain score (Hedges, 1981)

$$ES_{gain} = (Xgain_t\text{-}Xgain_c/sd_{pooledgain})$$

ES_{gain} is the standardized gain score effect size
$Xgain_t$ is the gain score of treatment group from pretest to posttest
$Xgain_c$ is the gain score of control group from pretest to posttest
$sd_{pooledgain}$ is the pooled standard deviation of gain score from pretest to posttest of treatment and control group

(3) Calculation of effect size using pre/post change score (Becker, 1988)

$$ES_{pp} = (X_{t2} - X_{t1}/sd_{pooled})$$
$$sd_{pooled} = ((S_{t1}^{\ 2} + S_{t2}^{\ 2})/2)^{1/2}$$

ES_{pp} is the standardized pre/post effect size
X_{t2} is the mean of treatment group in posttest
X_{t1} is the mean of treatment group in pretest
sd_{pooled} is the pooled standard deviation of pretest and posttest scores of treatment group
S_{t1} is the standard deviation of pretest score of treatment group
S_{t2} is the standard deviation of posttest score of treatment group

Appendix 2

Homogeneity Analyses and Mean Effect Size for Possible Moderators for Achievement

Table 1.*Participants*

Variable	Q_B	Q_W	Mean Effect Size
Education Level of Tutee	28.31***		0.65
Kindergarten		1.70***	0.42
Elementary (s1-s6)		323.80***	0.62
Middle school (s7-s9)		65.98***	1.01
Upper secondary (s10-s12)		14.73***	0.97
College or University		13.90***	0.21
Unspecified		1.45	0.63
Education Level of Tutor	29.12***		0.66
Kindergarten		1.70	0.42
Elementary (s1-s6)		311.34***	0.62
Middle school (s7-s9)		73.76***	1.02
Upper secondary (s10-s12)		14.89***	0.88
College or University		13.901***	0.21
Unspecified		1.14	0.70
Age of Tutee	14.49**		0.65
5-12 years		311.41***	0.63
13-18 years		59.58***	0.92
19+		17.47**	0.34
Unspecified		44.76***	0.74
Age of Tutor	17.01***		0.65
5-12 years		300.29***	0.63
13-18 years		69.73***	0.92
19+		17.47**	0.34
Unspecified		43.21***	0.65

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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

Table 1 (Continued)

Variable	Q_B	Q_W	Mean Effect Size
Socio-Economic Status (SES) of Participants	0.14		0.65
Low		160.84***	0.64
Mixed		103.54***	0.67
Unspecified		183.19	0.65
Ethnicity (Main proportion) of Participants	43.17***		0.64
Caucasian		123.96***	0.98
Afro-American		9.52	0.74
Black		3.42	0.42
Mixed		40.15***	0.61
Other		5.79	0.27
Unspecified		216.68***	0.75
Academic ability of Tutee	23.86***		0.65
Low		62.05***	0.96
Special need		58.82***	0.50
Mixed		70.17***	0.56
Unspecified		232.81***	0.74
Academic ability of Tutor	42.41***		0.65
Low		51.99***	1.18
High		0.17	0.24
Special need		46.65***	0.49
Mixed		68.28***	0.57
Unspecified		238.2***	0.73
Target Sample	23.77***		0.67
Selected Individuals		151.3***	0.87
Selected whole class		255.09***	0.55

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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

Table 2. *Methodology*

Variable	Q_B	Q_W	Mean Effect Size
Type of peer tutoring	10.81**		0.67
Same-age reciprocal peer tutoring		319.57***	0.74
Same-age non-reciprocal peer tutoring		60.79***	0.63
Cross-age non-reciprocal peer tutoring		48.71***	0.44
Involvement of teacher	0.31		0.65
Yes (teacher-led or monitor)		428.17***	0.64
No		19.23*	0.69
Structural Tutoring	16.01***		0.65
Yes		376.86***	0.72
No		48.87***	0.50
Mixed		5.97	0.26
Fidelity check	0.10		0.65
Yes		363.75**	0.65
No		*	0.65
		83.86***	
Control for author bias	65.84***		0.65
Yes (standardized test used)		172.11**	0.50
No (not standardized test used)		*	1.01
Mixed		209.59**	-0.02
		*	
		0.16	
Tutor Training	11.85***		0.65
Yes		401.39**	0.71
No		*	0.44
		34.46***	
Substitute to classroom instruction	56.52***		0.65
Yes		24.03	0.26
No		367.15**	0.79
		*	
Supplement to classroom interaction	28.99***		0.65
Yes		402.55**	0.73
No		*	0.28
		16.68	
Control and Treatment	3.43		0.46
Equivalent comparison group		0.76**	0.65
Non-equivalent comparison group of convenience		17.93	0.35
Random assignment of individuals to control and experimental groups		21.15	0.55
Random assignment of groups/classes/school to control and experimental groups		51.14*	0.45

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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

Table 3. *Intervention*

Variable	Q_B	Q_W	Mean Effect Size
Duration of tutor training	7.95***		0.57
(No. of sessions of tutor training)			
< 3 sessions		76.11***	0.65
≥ 4 sessions		11.71	0.35
Length of each session (minutes)	0.20		0.53
≤ 40 minutes		36.06***	0.51
> 40 minutes		39.66**	0.55
Duration of Intervention	12.21***		0.71
(No. of sessions per week)			
≤ 3 sessions		104.19***	0.51
≥ 3 sessions		219.13***	0.80
Length of each session of tutoring (minutes)	26.55***		0.69
≤ 30 minutes		272.04***	0.85
> 30 minutes		52.80***	0.49
Duration of tutoring (number of week)	1.69		0.66
≤ 12 weeks		139.68***	0.60
> 12 weeks		261.31***	0.69
Intervention setting	11.00		0.65
During school lesson		379.00***	0.67
Other		34.27**	0.70
Unspecified		20.34***	0.36

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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

Table 4. *Intervention Outcomes*

Variable	Q_B	Q_W	Mean Effect Size
Target Subject matter on achievement	9.02*		0.67
Mathematics		28.19*	0.52
Reading		55.94***	0.63
Other		322.21***	0.76
Quality of Scale for achievement on Target Subject matter			
1. Mathematics	0.86		0.49
Standardized measure		42.65***	0.52
Created for the study		13.36**	0.38
2. Reading	23.34***		0.55
Standardized measure		52.02**	0.50
Created for the study		0.01	1.48
3. Other Achievement Measure	37.47***		0.72
Standardized measure		111.42***	0.45
Created for the study		203.43***	0.98
Data Derived for Calculation of Effect Size	104.15***		0.65
Treatment-control posttest score		63.73**	0.33
Treatment-control gain score		21.22**	0.76
Pretest-posttest change score		258.60***	1.02

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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

Appendix 3

Homogeneity Analyses and Mean Effect Size for Self-concept

Variable	Q_B	Q_W	Mean Effect Size
Focus of Intervention	57.00***	97.22***	0.88
Target Self-concept		19.81***	1.09
Nontarget Self-concept		30.14	0.18

Note. A significant Q_B indicates a significant moderator whereas a non-significant Q_W shows that the variable can be grouped into homogenous subgroups.

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