School Psychology Review,
2013, Volume 42, No. 1, pp. 39–55

Academic Benefits of Peer Tutoring: A Meta-Analytic Review of Single-Case Research

Lisa Bowman-Perrott, Heather Davis, Kimberly Vannest, and Lauren Williams
Texas A&M University

Charles Greenwood
University of Kansas

Richard Parker
Texas A&M University

Abstract. Peer tutoring is an instructional strategy that involves students helping each other learn content through repetition of key concepts. This meta-analysis examined effects of peer tutoring across 26 single-case research experiments for 938 students in Grades 1–12. The TauU effect size for 195 phase contrasts was 0.75 with a confidence interval of CI_{95} = 0.71 to 0.78, indicating that moderate to large academic benefits can be attributed to peer tutoring. Five potential moderators of these effects were examined: dosage, grade level, reward, disability status, and content area. This is the first peer tutoring meta-analysis in nearly 30 years to examine outcomes for elementary and secondary students, and extends previous peer tutoring meta-analyses by examining disability as a potential moderator. Findings suggest that peer tutoring is an effective intervention regardless of dosage, grade level, or disability status. Among students with disabilities, those with emotional and behavioral disorders benefitted most. Implications are discussed.

The peer tutoring research base spans more than 40 years and convincingly demonstrates an evidence-based practice (Cloward, 1967; Cohen, Kulik, & Kulik, 1982; Delquadri, Greenwood, Whorton, Carta, & Hall, 1986; Mastropieri, Spencer, Scruggs, & Talbott, 2001). Peer tutoring can be defined as "a class of practices and strategies that employ peers as one-on-one teachers to provide individualized instruction, practice, repetition, and clarification of concepts" (Utley & Mortweet, 1997, p. 9). The success of peer tutoring for both tutors and tutees is likely from incorporated instructional features such as frequent opportunities to respond, increased time on task, and regular and immediate feedback. Each of these components is empirically linked with increased academic achievement

Correspondence regarding this article should be addressed to Lisa J. Bowman-Perrott, Texas A&M University, Department of Educational Psychology, 4225 TAMU, College Station, TX 77843-4225; e-mail: lbperrott@tamu.edu

(Greenwood, Terry, Arreaga-Mayer, & Finney, 1992; Maheady, Harper, & Sacca, 1988).

The positive effects of peer tutoring have been demonstrated across subjects such as reading (Oddo, Barnett, Hawkins, & Musti-Rao, 2010), math (Hawkins, Musti-Rao, Hughes, Berry, & McGuire, 2009), social studies (Lo & Cartledge, 2004), and science (Bowman-Perrott, Greenwood, & Tapia, 2007), and across a wide range of settings that include general education classrooms (Lo & Cartledge, 2004), resource rooms (Maheady et al., 1988), self-contained classrooms (Sutherland & Snyder, 2007), alternative placements (Bowman-Perrott et al., 2007), and group homes (Mayfield & Vollmer, 2007). Peer tutoring configurations include cross-age (Jun, Ramirez, & Cumming, 2010), small group (Maheady, Sacca, & Harper, 1987), and classwide (Greenwood et al., 1992). In addition, peer tutoring is effective for students with and without disabilities, native English-speaking students, and English language learners (see Okilwa & Shelby, 2010).

Although previous peer tutoring research indicates that student outcomes are better with the use of peer tutoring (Delquadri et al., 1986), there are some gaps in the literature. Missing from the peer tutoring literature are recent reviews that report effect sizes (ES) with confidence intervals for elementary and secondary students. Further, potential moderators have not been fully examined, and an evaluation of single-case data using a common effect size metric is needed.

Single-Case Research, Effect Size, and Confidence Intervals

Single-case research methods can "provide a rigorous experimental evaluation" of the efficacy of an intervention (Kratochwill et al., 2010, p. 2). As such, single-case research has been used to identify a range of interventions used in schools, as this method of inquiry can help identify practices that are evidence-based (Horner et al., 2005). The use of effect size in single-case research allows for a determination of the size or magnitude of academic or behavioral change. Determining the size of the effect, as well as a functional relation, is critical in light of accountability for instructional practices and multitier models of early intervention (see Council for Exceptional Children, 2008; National Association of School Psychologists, 2010).

Data from single-case studies of school-based practices are being summarized more as new methods are being developed that can address positive baseline trends and that require few assumptions about the data (Parker, Vannest, Davis, & Sauber, 2011). Although many studies using single-case research designs may be found in the peer tutoring literature, neither individual nor aggregated effect sizes with corresponding confidence intervals have been published to date. This is a significant shortcoming, as effect sizes aid in summarizing data across studies. Further, confidence intervals are needed for accurate interpretation of effect size data (Cooper, 2011; Hunter et al., 1982; Thompson, 2002, 2007) and are required by the American Psychological Association (APA; American Psychological Association, 2010; Wilkinson & APA Task Force on Statistical Inference, 1999). An effect size with confidence intervals tells about the relative size of an effect compared to other treatments, and provides a standard metric for comparison and aggregation. Further, in an era of evidence-based practices, it provides data that are more readily understood. The use of a common effect size metric with single-case research, as with group designs, is essential to allow for the aggregation of results across studies.

Effect Size Metrics Used in Previous Single-Case Research

There are at least eight commonly used nonoverlap effect size metrics (indices) in single-case research. They include percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), percentage of all nonoverlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), nonoverlap of all pairs (NAP; Parker, Vannest, & Brown, 2009), extended celeration line (ECL; White & Haring, 1980), improvement rate difference (IRD;

Parker et al., 2009), percent of data exceeding the median (PEM; Ma, 2006), Pearson's Phi (Parker et al., 2007), and Kendall's TauU for nonoverlap between groups with baseline trend control (Tau_{novlap}; Parker, Vannest, Davis, & Sauber, 2011). Although a full review of these indices is beyond the scope of this article (see Parker et al., 2011), a brief review of limitations of each compared to TauU follows.

Although not intended by its originators as an effect size, PND has been used this way in several meta-analyses. Limitations of PND include its unknown distribution qualities (with the consequent inability to provide a standard error or confidence intervals) and its insensitivity to treatment effects. As such, it is not recommended for meta-analyses (see Allison & Gorman, 1993; Horner et al., 2005). A limitation of PND, PAND, NAP, IRD, PEM, and Phi, is that they cannot control for positive baseline trend. Of the eight methods, only ECL and TauU can do that. However, ECL controls for only linear trend, whereas TauU controls for any shape of increase, known as monotonic trend. The weakest statistical power among these eight indices is shown by PEM, followed by ECL (it cannot be ascertained with PND). Medium statistical power is obtained by Phi and IRD; strongest statistical power is by TauU and NAP. Limitations of ECL and PEM also include their assumption that the median value is a reliable summary of Phase A. This assumption is only correct for data sets demonstrating measures of central tendency. TauU does not make this assumption.

Another criterion for a good effect size is that it should discriminate among results from different studies. A nondiscriminating index will lump all low and/or all high results together. Worst discrimination is shown by PEM, followed by PND. Best discrimination is by TauU, with baseline trend control. Phi also has good discriminating ability. TauU is also robust to autocorrelation (with little impact on standard error; Parker et al., 2011). Finally, TauU is well suited to very short phases. It does not require the minimum four to six expected data points per cell that nonparametric methods based on cross-tabulation do (e.g., IRD, Phi).

The limitations of TauU include it being a relatively new effect size measure, so there are relatively few publications from which its sizes can be judged. A second limitation is shared by all nonoverlap indices but Phi: nonoverlap effect sizes are not directly comparable to the familiar Pearson's *r* or Cohen's (1988) *d*. However, TauU's strengths (statistical power, low *N* requirement, distribution-free, good discrimination ability among studies, control of unwanted baseline trend, a known sampling distribution) are such that it appears to be the best nonparametric index at present and is even competitive with parametric indices.

Effect Size Measures Used in Previous Peer Tutoring Meta-analyses of Group Design Studies

Although effect sizes used in group design meta-analyses are not directly comparable to TauU, a brief review of their limitations with regard to single-case data are presented. Previous peer tutoring meta-analyses of group design studies used Cohen's *d*, Hedges' *g*, *t,* and *F* tests to calculate effect sizes. The most common effect size measure was Cohen's *d*. An advantage of Cohen's *d* and Hedges' *g* is that they are scale free or standardized. The main limitation of both is that they are parametric statistics. As such, they make assumptions about single-case data that are rarely all met. Assumptions include normality, constant variance along the time dimension, and an interval-level scale (see Hunter et al., 1982). In addition, parametric effect size measures permit results to be unduly influenced by a few extreme scores. Such problems are not inherent in the effect size measure but in the mismatch between parametric calculations and single-case data. The problem of unmet assumptions applies to all effect size measures based on parametric analyses (e.g., *t* tests, *F* tests, and linear regression); these measures are not sensitive to data trend in any phase. A review of effect sizes reported in previous meta-analyses follows.

Previous Meta-analyses of Peer Tutoring and Potential Moderators

Cohen et al. (1982) examined school tutoring programs across 65 studies that included students in Grades 1–12. Variables included grade level, duration of the intervention (number of weeks), and content area. Overall effect sizes were reported for tutors and tutees separately, 0.33 and 0.40, respectively. Students who participated in tutoring outperformed students in control groups on content area tests (identified as math, reading, and "other"). Further, elementary age students benefitted more than their older (middle and high school) peers, and engagement for fewer weeks yielded a larger effect size (0.95) than engagement for a longer period of time (ES = 0.42 for 5–18 weeks and ES = 0.16 for 19–36 weeks).

Although Cohen et al. (1982) focused on students without disabilities, Cook, Scruggs, Mastropieri, and Casto (1985) examined 19 studies of peer tutoring arrangements in which elementary and secondary students with disabilities served as tutors and tutees. Tutors were identified as students with emotional/behavioral disorders (EBD), learning disabilities (LD), and mental retardation (MR) or intellectually disability (ID); the majority consisted of students with EBD (56%). Nearly 40% of the tutees were students with LD (20%) and MR (18%). Both tutors and tutees made academic gains as a result of participating in peer tutoring (ES = 0.59 and 0.65, respectively). Cook et al. (1985) evaluated the age of tutors and tutees, the number of tutoring sessions per week and total number of sessions, the number of hours of tutoring, and the length of sessions (minutes). Unlike the results of Cohen et al.'s (1982) analysis, the length of the intervention (treatment dosage) did not appear to influence the outcomes.

Nearly two decades later, Rohrbeck, Ginsburg-Block, Fantuzzo, and Miller (2003) examined 90 studies of peer-assisted learning (PAL) interventions at the elementary school level. In this investigation, larger gains were found for students in Grades 1–3 (ES = 0.37) than for students in Grades 4–6 (ES = 0.28). Further, larger gains were found when reward contingencies were present (ES = 0.34 vs. 0.26). Like the meta-analysis by Cook et al. (1985), but unlike that of Cohen et al. (1982), dosage (the total number of hours or minutes in which participants participated in peer tutoring) did not affect academic achievement.

Ginsburg-Block, Rohrbeck, and Fantuzzo (2006) examined PAL interventions across 36 studies involving elementary school students. No potential moderator variables of academic achievement were assessed. However, they reported an effect size of 0.48 for academics for the 26 studies that reported those data. Finally, Jun et al. (2010) analyzed 12 studies to examine the impact of peer tutoring on literacy outcomes for students in Grades 6–12. Effect sizes were compared across types of peer tutoring (e.g., cross-age peer tutoring vs. adult tutoring). The effect size for cross-age peer tutoring was 1.05.

Existing meta-analyses provide valuable information about peer tutoring, but with some limitations. The meta-analysis conducted by Cohen et al. (1982) did not report fidelity of implementation. In addition, effect sizes were only reported for reading, math, and unidentified subjects in a category labeled "other." Similarly, Cook et al.'s (1985) meta-analysis did not report treatment fidelity. Although the mean age of tutors and tutees were elementary and middle school, some tutors were adults, but age or grade as a moderator of effects was not evaluated. Also, the effect of disability as a potential moderator was not examined and data were not disaggregated by disability category. Although Rohrbeck et al. (2003) reported effect sizes for several content areas subjects, the findings for the studies examined in their meta-analysis were limited to elementary school students. Like Rohrbeck et al. (2003), the Ginsburg-Block et al. (2006) meta-analysis findings were limited to elementary-age students. Although an effect size for academics was computed to correlate with students' social and behavioral outcomes, academic achievement was not examined. Jun et al.'s (2010) meta-analysis did not examine intervention variables or potential moderators

related to peer tutoring (e.g., age of participant or treatment dosage). Further, finds were limited to secondary students and literacy outcomes. Although effect sizes were reported in each of the five existing meta-analyses, only the three most recent reported confidence intervals for the effect sizes. Of those three, only two focused on academic outcomes (Jun et al., 2010; Rohrbeck et al., 2003). Finally, none of the five existing meta-analyses included studies using single-case research designs.

The Need for a Single-Case Research Meta-analysis

Studies utilizing single-case research designs have often been excluded from meta-analyses (Allison & Gorman, 1993; Cumming, 2012). The peer tutoring literature is no exception. It is likely that these studies have been excluded because standards for high-quality single-case research designs and evidence of treatment effects have been developed only recently (e.g., Cooper, 2011; Horner et al., 2005; Horner & Kratochwill, 2012; Kratochwill et al., 2010). Thus, methodology for standardizing, aggregating, and analyzing single-case data now provide an opportunity for their potential contribution to the literature.

The present meta-analysis examined the effect of peer tutoring and potential moderators on academic achievement for elementary and secondary students. As such, it was designed to contribute to the peer tutoring literature by extending previous research in several ways. Specifically, this meta-analysis is (a) the first peer tutoring meta-analysis in nearly 30 years that reports peer tutoring effects for both elementary and secondary students; (b) the most recent analysis of peer tutoring studies; (c) the only peer tutoring meta-analysis that examines disability as a potential moderator; (d) the first peer tutoring meta-analysis to investigate the contribution of studies using single-case designs by proposing an effect size metric that allows for the aggregation of single-case data; (e) the first single-case peer tutoring meta-analysis to provide support for peer tutoring as an evidence-based practice based on APA standards of reporting effect sizes with confidence intervals; and (f) the only peer tutoring meta-analysis to use interobserver agreement as an inclusion criterion. Two research questions were addressed: (a) What is the effect size of peer tutoring across studies? (b) What are examined effects of potential moderators on students' academic achievement?

Method

To identify relevant studies published in peer-reviewed journals, a review of the literature of data-based peer tutoring studies was conducted using the Education Full Text, Educational Resources Information Center (ERIC), and PsycINFO databases. Google Scholar was searched as well. Search terms included "peer tutoring," "reciprocal peer tutoring," "classwide peer tutoring," "peers as tutors," "peer-mediated instruction," and "peer-assisted learning" to identify as many peer tutoring studies as possible. A total of 1,758 articles were found. An ancestral search was conducted for studies in the reference section of articles using single-case designs. An additional ancestral search was completed for studies included in peer tutoring literature reviews. These uncovered 17 articles. These searches yielded a total of 1,775 articles.

Inclusion Criteria, Quality of Research Design, and Experimental Control

Several criteria had to be met for articles to be included in these analyses. First, studies had to be published in peer-reviewed journal articles between 1966 and 2011. Second, studies had to employ a single-case research design with baseline conditions that did not involve some form of peer tutoring. Third, research designs had to meet the criteria for strong single-case designs (Horner et al., 2005; Horner et al., 2012; Kratochwill et al., 2010). Specifically, (a) the peer tutoring intervention had to be systematically manipulated; (b) academic achievement outcome variables had to be measured using interobserver agreement of at least 80% for at least 20% of all observations; (c) studies had to demonstrate

experimental control by at least three demonstrations of the effect of the intervention at three points in time, and (d) phases had to have a minimum of three data points. The criteria established by Horner et al. (2012) for determining a functional relation between independent and dependent variables were also applied. Thus, studies with designs that did not meet these criteria (e.g., AB, ABA designs) were excluded. In addition, studies had to include the use of peer tutoring as an academic intervention in a content area. The term "content area " is used to reflect the core academic subject areas specified in the Individuals with Disabilities Education Act (2004; e.g., reading, math, social studies). Studies also had to use a direct measure of academic achievement. Finally, studies had to include students in Grades 1–12 involving same- or cross-age peers as tutors.

The following were excluded from the current meta-analysis: duplicate studies (i.e., studies that appeared in more than one database search); studies involving tutoring between college students; the use of college students, parents, or other adults as tutors for students in Grades 1–12; studies that investigated the use of peer tutoring in subject areas not identified in Individuals with Disabilities Education Act (2004; e.g., physical education); and reviews of literature and descriptive reviews of peer tutoring programs. Group design studies were excluded because previously published meta-analyses examined studies using these designs. A total of 26 articles remained.

Coding Reliability and Intercoder Reliability

Reliability was calculated for 20% (n = 39) of all of the AB phase contrasts across the 25 studies for which graphed or raw data were provided. Three data columns were used for each data series: phase, time, and score. The formula used for intercoder reliability was the sum of agreement/total number of agreements + disagreements × 100. Reliability was 100% for each code as described below. Initial agreement across variables ranged from 86% to 100%. Disagreements were resolved after the first author and graduate student reread and discussed articles, resulting in 100% final agreement across all codes. The coding guide is available from the first author upon request.

Codes were operationally defined and applied to all 26 studies by the first author; all codes were entered into an Excel spreadsheet. A graduate student was trained on the codes and independently coded each study using a separate Excel spreadsheet. Thus, each study was double-coded. Twelve study variables were coded, including the five potential moderators. Coding across potential moderators included (a) grade level; (b) dosage (including its components of duration, intensity, and number of sessions); (c) use of reward; (d) disability or at-risk status; and (e) design strength/experimental control. Reliability was calculated for each of the three components of dosage separately, and then for the dosage formula. Five of the studies did not provide some of the necessary coding information for the potential moderators (e.g., length of the intervention). In these cases, the first or second author was contacted via e-mail to obtain the missing data; all of the authors responded and provided the requested information.

In addition to the potential moderators, all studies were coded for design strength and fidelity of implementation. Reliability was calculated for the five moderator variables, design strength, fidelity of implementation, and 20% of the remaining variables (e.g., participant characteristics such as gender, study characteristics such as number of participants, and social validity). Twenty-five percent of all codes were calibrated. As a result, several codes were changed or deleted. Duration, intensity, and number of sessions were initially coded separately. They were subsequently combined to create one variable: dosage. Peer tutoring format (viz., use of reciprocal tutoring, error correction, and awarding of points) was not retained.

Potential Moderators

Studies were coded across five potential moderator variables: grade level, dosage, the

use of rewards, disability or at-risk status, and content area. Potential moderators were coded by level with the exception of content area, which was simply coded by subject.

Grade level. Grade level was represented by two levels: elementary and secondary. Elementary grades included Grades 1–5. Secondary grades represented middle and high school Grades 6–12.

Dosage. Dosage was composed of the variables duration, intensity, and number of sessions. The formula used to calculate this potential moderator was duration × intensity × number of sessions (see Rohrbeck et al., 2003). Duration refers to the number of weeks students spent involved in peer tutoring, rather than the entire length of the study (which includes baseline or a nonpeer tutoring condition). Intensity refers to the number of minutes students spent engaged in peer tutoring (per week). Number of sessions refers to number of times (e.g., days per week) students engaged in the intervention across the weeks of the peer tutoring intervention.

Use of rewards. Tangible (e.g., stickers) and social (e.g., applause) rewards were represented across studies. The use of individual contingencies alone or with group contingencies varied across studies.

Disability or at-risk status. Disability status referred to students who had been identified with a disability under Individuals with Disabilities Education Act. At-risk status was applied to students who performed below grade level or who were performing poorly academically, but were not identified as having a disability.

Content area. Content areas included reading, math, and social studies. Ten studies focused on reading, six on spelling, six on math, three on vocabulary, and three on social studies (some studies addressed more than one content area). Spelling outcomes were measured by students' correct spelling of words and the percent of words capitalized correctly. Reading measures included (a) nonsense word fluency, (b) errors per minute, (c) sight word acquisition, and (d) comprehension. Math measures consisted of (a) correctly multiplying decimals, (b) changing decimals to fractions, (c) calculating percentages, and (d) adding and subtracting time. Vocabulary evaluated the percent of vocabulary words correct. Social studies included history content material; specific learning outcomes were not reported.

Calculation of Effect Size

TauU. TauU is an effect size measure based on nonoverlap between phases that can also control for confounding baseline trends (Parker et al., 2011). It is derived from Kendall's rank correlation (an index of trend) and the Mann-Whitney *U* test between groups (an index of nonoverlap; Parker et al., 2011). Kendall's rank correlation is essentially an analysis algorithm of time and score, comparing ordered scores and all possible pairs of data. Each pairwise comparison is an improved score, a score that is not improved, or a tie. Kendall's rank correlation is the percentage of all data pairs that show improvement and that measures the tendency for scores to improve over time. It also calculates monotonic trendedness. In Mann-Whitney *U*, the index represents differences in group level. In application to single-case research, the concept is applied to phases (rather than groups). Scores from two phases (groups) are combined for a cross-group ranking. The separate rankings are then statistically compared for mean differences. The Mann-Whitney *U* algorithm uses two continuous variables: scores and time. By replacing the time variable with a dummy code (0/1) to represent Phases A and B, an identical result is produced. This produces the proportion of pairwise comparisons that improve from Phase A to Phase B or the percentage of nonoverlapping data.

TauU and phase contrasts. TauU incorporates A versus B phase nonoverlap, nonoverlap and Phase B trend together, nonoverlap with baseline trend controlled, and nonoverlap and Phase B trend with baseline trend controlled. It is constrained by the amount of

Phase A trend, Phase A length, and the relative lengths of Phase A and B. Therefore, it can only control baseline trendedness to the rational limits. In the present meta-analysis, data from all AB phases were analyzed, with the exception of maintenance phases. In studies in which more than one AB phase was reported, an effect size was calculated for the A1/B1 contrast and a separate effect size for the A2/B2 contrast. Each effect size was then entered into WinPepi (Abramson, 2012) using the meta-analysis function to aggregate the data and arrive at a single effect size for a given study. Specifically, the following data analysis functions were selected: (a) *Compare 2*, (b) *Meta-analysis; analysis of stratified data*, (c) *Others, or proportions or rates with effect sizes/CIs*, and (d) *Also enter standard error*. Each AB contrast, or "stratum," was entered, and all were "combined" (Abramson, 2012).

TauU dummy coding. Dummy coding for A and B phases (e.g., A1/B1, A2/B2) were calculated by hand for each study and entered into the TauU calculator (Vannest, Parker, & Gonon, 2011) to obtain values for Tau (herein used synonymously with TauU) and the standard error of Tau (SE_{Tau}). Tau and SE_{Tau} values were entered into WinPepi to arrive at an effect size and confidence interval for each study. Because one article contained two studies (Greenwood et al., 1984), the number of articles and the number of studies were not equal in all analyses.

Cohen's *d* and Tau transformation. In one study (Kamps et al., 2008), raw scores were not available; means and standard deviations were reported. For this study, Cohen's *d* was first calculated and then transformed to TauU (Rosenthal, 1994). In this instance, TauU was calculated by using a three-step process. First, a Cohen's *d* effect size was calculated by hand using the formula Cohen's $d = (M_1 - M_2)/\sigma_{pooled}$ where $\sigma pooled = \sqrt{[(\sigma 1^2 + \sigma 2^2) / 2]}$. The Cohen's *d* effect size was also obtained from WinPepi along with Cohen's *d SE* (Abramson, 2011). Second, Cohen's *d* was transformed to Tau using the formula $Tau = 1 - (1 - d/3.464)^2$ (Acion, Peterson, Temple, & Amdt, 2006; Parker, Vannest,& Brown, 2009). Third, SE_{Tau} was calculated from the Cohen's *d* standard error by dividing Cohen's *d* by Tau and then dividing the Cohen's *d* value by the product. For example, dividing a Cohen's *d* of 1.39 by a Tau of 0.64 yields a quotient of 2. Dividing the Cohen's *d* standard error of 0.57 by 2 yields a quotient of 0.28, the SE_{Tau}. Transformed Cohen's *d* values were entered into WinPepi.

Statistical significance. Statistical significance for Tau values was determined using confidence interval CI_{95}. When determining wheher change is reliable, a 90%–95% confidence interval is standard (Nunnally & Bernstein, 1994), indicating a reasonable change of 5–10% likelihood of error. Statistical significance between Tau values was determined by calculating $CI_{83.4}$ to visually test for overlap of upper and lower limits between effect sizes. Visual comparison of two effect sizes with $CI_{83.4}$ is the same as a $p = .05$, or 95% confidence-level test between the two scores (Goldstein & Healy, 1995; Payton, Greenstone, & Schenker, 2003).

Results

Study Characteristics

The 26 studies examined in this meta-analysis were published between 1984 and 2011 and included a total of 938 participants. Although data on participant gender were not reported in four of the studies, the majority of participants were reported as being male. Fourteen studies did not report participant ethnicity. Among those that did report ethnicity, Caucasian and African American were the two ethnic groups with the greatest representation. Studies largely involved students with identified disabilities ($n = 15$) and/or those at risk for disabilities ($n = 12$). Only four studies did not report the inclusion of students with or at risk for disabilities. Most of the studies ($n = 17$) were implemented in general education classrooms, followed by special education classrooms (e.g., self-contained, resource). One study was implemented in an English-as-

a-second language classroom; another was implemented in a group home and private homes.

The most commonly used design was a multiple-baseline design ($n = 15$), with three or more replications of the independent variable across at least three points in time. Two studies used a multiple-probe design (also demonstrating at least three replications across three points in time); six used an ABAB design. One study used an alternating-treatment design comparing a peer tutoring and "no peer tutoring" condition. Cases represented individual student data in 14 studies, classes in 10 studies, and peer tutoring dyads in two studies.

Interobserver agreement was reported in all 26 studies for at least 20% of all observations with 80% or more agreement (average agreement is 97.69%) (Kratochwill et al., 2010). Treatment fidelity data were reported in 16 of the studies, ranging from 82.86% to 100%. Mean reliability for student implementation was 93.64%; reliability was 94% for teachers. The majority of studies did not collect consumer satisfaction (social validity) data ($n = 11$). Those that did ($n = 7$), collected teacher and student feedback via questionnaires or surveys; one collected data from parents. Satisfaction ratings were high among teachers, students, and parents.

Overall Effect

In response to the first research question, the overall effect of peer tutoring was examined across all 26 studies, yielding a mean effect size of 0.75 ($SE = 0.02$, $CI_{95} = 0.71$ to 0.78). Figure 1 illustrates the range of effect sizes and confidence intervals across all of the studies at a 95% confidence level. Thus, there is a 95% certainty that the true value for the obtained effect size fell between the upper and lower limits of the calculated confidence interval.

Findings for Potential Moderators

Potential moderators of peer tutoring were tested by calculating a reliable difference using the formula (L1–L2)/sqrt [(*SE*Tau1Sqrd) + (*SE*Tau2Sqrd)]. Results indicated whether levels of each moderator reliably reflected differences in the effect of peer tutoring (see Table 1). If statistically significant differences were obtained between levels, the potential moderator was confirmed as a moderator because they differentially affected students' outcomes. Potential moderators were coded by level with the exception of content area, which was simply coded by the subject areas represented across studies. Results address the second research question.

Grade level. Peer tutoring was found to be a slightly more effective intervention for middle and high school students (ES = 0.74, $SE = 0.04$, $CI_{95} = 0.66$ to 0.81) than for elementary school students (ES = 0.69 $SE = 0.02$, $CI_{95} = 0.65$ to 0.74), as indicated by the overlap at CI_{95} (see Table 2). The reliable difference results for grade level were $z = 1.11$, $p = .263$.

Dosage. The median dosage was 480 min with an Inter Quartile Range of 280 to 1,137.5 min. Of the dosage amounts reported, 75% ($n = 13$) were greater than 1,137.5 min and 25% ($n = 12$) were greater than 280 min. One of the studies (Greenwood et al., 1984) had an Inter Quartile Range of 480. Studies with a dosage value below the median had the same effect size (.75, $SE = 0.03$, $CI_{95} = 0.69$ to 0.81) as those with values at or above the median (ES = 0.75, $SE = 0.02$, $CI_{95} = 0.70$ to 0.79). The obtained reliable difference values were $z = 0$, $p = 1$.

Use of rewards. Studies ($n = 13$) that employed the use of rewards (e.g., rewards vs. no rewards) had a higher ES (.75, $SE = 0.02$, $CI_{95} = 0.71$ to 0.79) than those that did not (ES = 0.69, $SE = 0.03$, $CI_{95} = 0.63$ to 0.73; $n = 13$). Further analysis by grade level revealed that middle and high school students (ES = 0.83, $SE = 0.08$, $CI_{95} = 0.68$ to 0.98) benefit from the use of rewards more than elementary school students (ES = 0.70, $SE = 0.03$, $CI_{95} = 0.65$ to 0.75). Reliable difference values were $z = 4.44$, $p = .001$.

Disability/at-risk status. Studies involving students identified with or at risk for

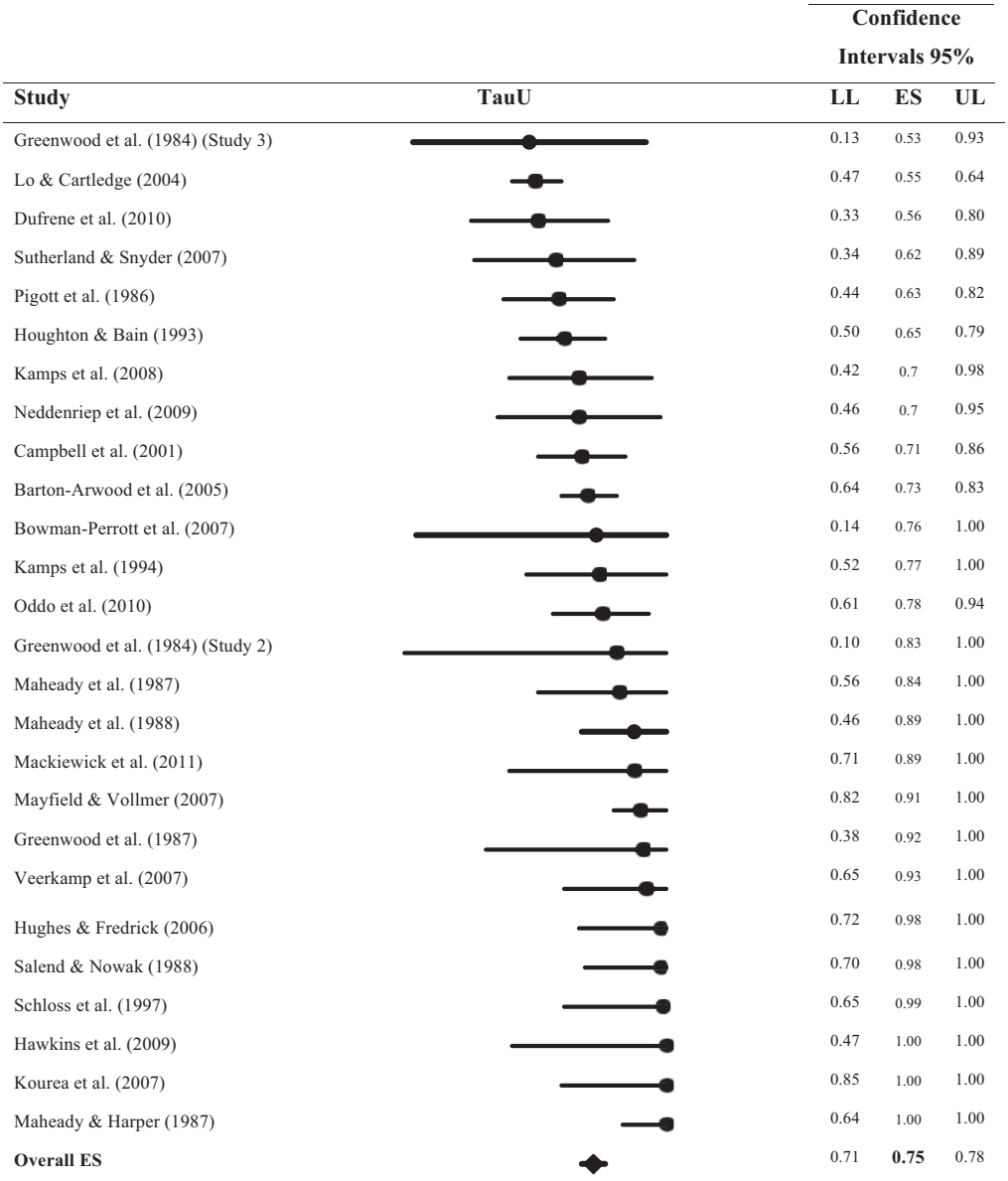

Figure 1. Forest plot for the effects of peer tutoring on students' academic outcomes. Circles represent effect sizes from individual studies. They are proportionate to their weight in the overall effect size. Circles and confidence interval bars represent the precision of each study effect size. The diamond represents the overall effect across all 26 studies. UL = upper limit; LL = lower limit; ES = effect size.

disabilities had an effect size of 0.76 (CI_{95} = 0.72 to 0.79). By comparison, studies that did not involve students with or at risk for disabilities had an effect size of 0.65 (CI_{95} = 0.51 to 0.79). The effect size for students with LD and EBD was 0.75 and 0.76, respec-

Table 1
Summary of Effect Size Results for Moderator Variables

Moderator	*k* (Studies)	*n* (Participants)	Mean Effect Size	Standard Error	95% CI	
					Lower Limit	Upper Limit
Grade level						
Elementary	12	136	.69	.02	.65	.74
Secondary	10	368	.74	.04	.66	.81
Dosage						
≤480 minutes	14	446	.75	.02	.70	.79
>480 minutes	12	209	.75	.03	.69	.81
Reward						
Yes	13	570	.75	.02	.71	.79
No	13	86	.69	.03	.63	.75
Disability						
Yes	23	511	.76	.02	.72	.79
No	4	28	.65	.07	.51	.79

Table 2
Reliable Difference Between Moderator Levels

Moderator	Effect Size	Standard Error	*z*	*p*
Grade				
Elementary	.69	.02		
Secondary	.74	.04	1.11	.263
Dosage				
≥480 min	.75	.03		
<480 min	.75	.03	0.00	1.00
Reward				
Yes	.72	.02		
No	.65	.03	4.44*	.001

Note. *$p < .001$, two-tailed test.

tively. The calculated reliable difference values for this variable were $z = 1.51$, $p = .935$.

Content area. Vocabulary yielded a large effect (ES = 0.92; $SE = 0.07$, $CI_{95} = 0.77$ to 1.00), followed by math (ES = 0.86; $SE = 0.04$, $CI_{95} = 0.78$ to 0.94), reading with a large to moderate effect size (ES = 0.77; $SE = 0.03$, $CI_{95} = 0.71$ to 0.82), spelling (ES = 0.74; $SE = 0.06$, $CI_{95} = 0.62$ to 0.85), and social studies a small effect size (ES = 0.57, $SE = 0.04$, $CI_{95} = 0.50$ to 0.65).

Discussion

This meta-analysis is the first peer tutoring meta-analysis to examine achievement outcomes for elementary and secondary students across peer tutoring studies using single-case research designs. The overall effect size was found to be moderately large, indicating that greater academic gains were achieved by students engaged in peer tutoring interventions than nonpeer tutoring instructional arrangements.

Moderator analyses revealed a statistically significant effect for the use of rewards. Peer tutoring interventions that used rewards had a larger effect size than those that did not. This finding points to the importance of the use of reward on academic outcomes, especially for middle and high school students. Findings from several peer tutoring studies support its use with older students as a means of motivation. This seems to be the case particularly for students who have experienced academic difficulties (see Bowman-Perrott et al., 2007; Kamps et al., 2008; Mitchem, Young, West, & Benyo, 2001). This finding is also consistent with previous research at the

elementary school level (e.g., Rohrbeck et al., 2003), as the use of reward contingencies produced a statistically significant effect.

In four studies, data for elementary and secondary students were not disaggregated. The two studies from Greenwood et al. (1984) reported combined data for students in Grades 3–6, Mayfield and Vollmer (2007) presented data for students in Grades 3–11, and the spelling class data from Bowman-Perrott et al. (2007) were reported for students in Grades 5–12 together. Mayfield and Vollmer implemented peer tutoring in group homes and students' homes. Bowman-Perrott et al. (2007) conducted peer tutoring in an alternative school that was part of a residential treatment facility. In these instances, the way students were grouped in their natural learning environments did not follow traditional grade-level groupings. Data for these studies were not included in grade-level analyses because they would not provide insight into the effect of peer tutoring on students at the elementary versus secondary level. Results demonstrated that peer tutoring was effective for both elementary and secondary students, and that grade level did not moderate its effectiveness. Although individual grade level analyses could not be conducted because data were not disaggregated by grade level in the majority of the studies, participants tended to represent certain grade levels. For studies involving elementary school students, the average grade level was fourth grade, followed by third grade. Studies focusing on secondary students tended to include sixth-graders most often, followed by ninth-graders.

Consistent with the Rohrbeck et al. (2003) study, the findings of the current meta-analysis showed no difference in student outcomes for studies with dosage amounts above and below the median value. Perhaps the core components of peer tutoring (e.g., increased opportunities to respond, error correction procedures) are sufficient to make an impact on student outcomes with as few as 280 min of exposure to the intervention and as many as just over 1,000 min. The finding that students with or at risk for disabilities demonstrated greater academic gains than their peers without disabilities or at-risk status may be reflective of the benefit students received from the additional support (e.g., more opportunities to respond) afforded by peer tutoring. This may be especially likely, as all of these students were identified as being below grade level in a given content area.

Twenty-three of the 26 studies included participants with identified disabilities or who were determined to be at risk for being identified as having a disability because of poor academic performance. Although differences were not statistically significant, practical significance can be attributed to the finding that the effect size was larger for students with or at risk for disabilities (.76) than for students without disabilities or who were not at risk (.65). Results support evidence that aspects of peer tutoring interventions such as repetition of key concepts and opportunities to respond are particularly beneficial for students in need of additional academic supports. Of the 23 studies that included students with or at risk for disabilities, only 11 disaggregated achievement outcomes by disability category. With regard to disability status, only data for students identified as having a LD or EBD as their primary disability were analyzed, as only one study disaggregated data for students with autism, one for students with MR, and one for students with other health impairments. It is important to note, however, that the number of studies for analyses of students with LD and EBD were small; caution should be used in considering these results.

Ten studies focused on reading, six on spelling, six on math, three on vocabulary, and three on social studies. Spelling outcomes were measured by students' correct spelling of words and the percentage of words capitalized correctly. Reading measures included (a) nonsense word fluency, (b) errors per minute, (c) sight word acquisition, and (d) comprehension. Math measures consisted of (a) correctly multiplying decimals, (b) changing decimals to fractions, (c) calculating percentages, and (d) adding and subtracting time. Vocabulary included the percent of vocabulary words correct. Finally, social studies included history; specific learning outcomes were not reported.

In the present analysis, reading yielded a large to moderate effect size (ES = 0.77), compared with the effect sizes reported by Cohen et al. (1982) of 0.29, Cook et al. (1985) of 0.30 for tutors and 0.49 for tutees, and Rohrbeck et al. (2003) at 0.26 for reading. The effect size obtained for math in this meta-analysis (ES = 0.86) was also larger than that reported by Cohen et al. (1982; ES = 0.60), Cook et al. (1985; ES = 0.67 and 0.85 for tutors and tutees, respectively), and Rohrbeck et al. (2003; ES = 0.22). Although social studies had a smaller effect in this meta-analysis (ES = 0.57), it was larger than that reported by Rohrbeck et al. (2003; ES = 0.49). The obtained moderate effect size for spelling (ES = 0.74) was larger than those reported by Cook et al. (ES = 0.01 and 0.51 for tutors and tutees, respectively) and Rohrbeck et al. (2003) (ES = 0.21). Finally, vocabulary had a large effect (ES = 0.92). Because previous meta-analysis reported data for writing, language, literacy, or a combination of related content areas, there is no vocabulary effect size with which to compare the present findings. The effect sizes for content areas should be considered with caution, as a small number of studies were available for analysis. This is particularly true for vocabulary and social studies.

As previously mentioned, treatment fidelity data were reported in 16 of the 26 studies (62%). Rohrbeck et al. (2003) reported that 68% of studies in their meta-analysis reported these data. It is important to consider the potential impact of treatment fidelity on study outcomes. It was not examined as a potential moderator because fidelity was high in the studies that reported it. Therefore, comparing studies with low fidelity to those with high fidelity was not possible. It is important to report these data to help understand the degree to which teachers and students accurately implement peer tutoring interventions.

Limitations

The findings of this meta-analysis should be considered in light of the following limitations. One limitation is the lack of disaggregated disability data in some studies, limiting our sample for these analyses. Similarly, data were not disaggregated by grade level in most of the articles, so could not present results and recommendations by grade level. Another limitation is the potential variability that was introduced by how academic gains were measured across studies (e.g., curriculum-based measures vs. standardized tests) and peer tutoring type (e.g., cross-age vs. same age). A third potential limitation is that the effect sizes for the content areas may change with a larger pool of studies (especially for vocabulary and social studies). A final limitation is that caution should be used in comparing TauU to Cohen's *d* effect sizes. For example, the conversion from Cohen's *d* to TauU for the Kamps et al. (2008) study is an approximation. Future research can help address some of these limitations to further add to the peer tutoring literature.

Implications for Research

The findings of this meta-analysis underscore several recommendations that can inform future research. The first is the need to report treatment fidelity in peer tutoring studies. Knowing whether high versus low levels of fidelity promote greater academic gains would be beneficial in informing practice. The second is that treatment fidelity could serve as a moderator of student outcomes. This should be investigated by grade and across content areas. In addition, it would be helpful to addresses practical questions such as the following: (a) Is there a minimum number of hours or sessions needed for students to gain the most benefit from peer tutoring? (b) Does the type of academic outcome measurement (e.g., criterion-referenced vs. norm-referenced) moderate the effect of peer tutoring? (c) Does the comorbidity of LD and EBD affect student outcomes? (d) What do outcomes look like for students with autism and other disabilities? The last question is important, as data in these analyses were limited to students with LD and EBD because the number of cases for students with other disabilities (e.g., autism) was too small to investigate.

The peer tutoring literature would also benefit from studies that disaggregate data by grade (e.g., first grade) because the effects of peer tutoring, and moderators of those effects, may vary within each of the grades. Analyzing data by grade level (viz, elementary or secondary) would be beneficial. For example, it would be helpful to know whether first-graders may benefit more than fifth-graders. This would also prove useful in further analyzing potential moderators. For instance, the use of rewards may be significant with sixth-graders but not eighth-grade students.

Another recommendation is that future single-case peer tutoring studies should apply strong designs in keeping with recent standards (e.g., Kratochwill et al., 2010). Unfortunately, 11 single-case studies were excluded from these analyses because of weak designs ($n = 6$) and because interobserver agreement standards were not met ($n = 5$; Kratochwill et al., 2010). Among those excluded for these reasons were the two studies that focused on science in middle and high school classrooms. Thus, there is a need to investigate students' outcomes in this core content area. Further, more studies are needed across content areas, as the small number of studies in some of the content areas prevented a thorough analysis of students' academic outcomes.

Future research should examine the relation between academic and behavioral outcomes for students engaged in peer tutoring. This would be particularly useful in light of the finding that students with EBD benefitted most from peer tutoring. It would be interesting to examine the benefit students receive from peer tutoring with regard to behaviors. Finally, as a new effect size measure, this meta-analysis should be replicated using TauU as new single-case studies investigating peer tutoring are published. Given the many advantages of TauU, it holds great promise for aggregating single-case research data.

Implications for Practice

This meta-analysis provides evidence for the use of peer tutoring as an evidence-based instructional practice based on the most current, high-quality standards for single-case research. Social validity data across peer tutoring studies revealed that teachers find it easy to implement within their existing teaching routine and structure. As such, teacher training programs and in-service training for practicing teachers should include peer tutoring. This is particularly important in an era of increased accountability for implementing evidence-based practices and the implementation of multitiered early intervention supports.

Peer tutoring is an effective intervention for students with disabilities. This is especially noteworthy for students with EBD who have been consistently identified in the literature as performing below grade level, and for whom academic deficits are part of the federal definition and criteria for identification in Individuals with Disabilities Education Act (2004). Problem behaviors, a characteristic of students with EBD, adversely affect academic performance (see Individuals with Disabilities Education Act, 2004). Results showed that students with EBD, who by nature of their disability demonstrate problem behaviors, are most likely to benefit academically from a peer tutoring instructional format. It is an intervention that is highly recommended for their peers without disabilities as well.

Peer tutoring is effective in promoting academic gains across content areas, and is effective for elementary, middle, and high school students. The use of rewards appear to benefit older students as a motivator.

Footnotes

*References marked with an asterisk indicate studies included in the meta-analysis.

References

Abramson, J. H. (2011). WINPEPI updated: Computer programs for epidemiologists, and their teaching potential. *Epidemiologic Perspectives & Innovations, 8*(1). Available at http://www.epiperspectives.com/content/8/1/1

Abramson, J. H. (2012). *WinPepi programs Compare 2 manual* (Version 2.69). Available at http://www.biomedcentral.com/1742–5573/content/1/1/6

Acion, L., Peterson, J. J., Temple, S., & Amdt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine, 25,* 591–602.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31,* 62–631.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

*Barton-Arwood, S. M., Wehby, J. H., & Falk, K. B. (2005). Reading instruction for elementary age students with emotional and behavioral disorders: Academic and behavioral outcomes. *Exceptional Children, 72*(1), 7–27.

*Bowman-Perrott, L. J., Greenwood, C. R., & Tapia, Y. (2007). The efficacy of peer tutoring used in secondary alternative school classrooms with small teacher/pupil ratios and students with emotional and behavioral disorders. *Education and Treatment of Children, 30*(3), 65–87.

*Campbell, B. J., Brady, M. P., & Linehan, S. (1991). Effects of peer-mediated instruction on the acquisition and generalization of written capitalization skills. *Journal of Learning Disabilities, 24,* 6–14.

Cloward, R. D. (1967). Studies in tutoring. *Journal of Experimental Education, 36,* 14–25.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational 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. *The Journal of Special Education, 19,* 483–492.

Cooper, H. (2011). *Reporting research in psychology: How to meet journal article reporting standards* (6th ed.). Baltimore, MD: American Psychological Association and United Book Press.

Council for Exceptional Children. (2008). *Classifying the state of evidence for special education professional practices: CEC practice study manual.* Arlington, VA: Author.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Delquadri, J., Greenwood, C. R., Whorton, D., Carta, J. J., & Hall, R. V. (1986). Classwide peer tutoring. *Exceptional Children, 52,* 535–542.

*Dufrene, B. A., Reisener, C. D., Olmi, D. J., Zoder-Martell, K., McNutt, M. R., & Horn, D. R. (2010). Peer tutoring for reading fluency as a feasible and effective alternative in responsive to intervention systems. *Journal of Behavioral Education, 19,* 239–256.

Ginsburg-Block, M. D., Rohrbeck, C. A., & Fantuzzo, J. W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology, 98,* 732–749.

Goldstein, H., & Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158,* 175–177.

*Greenwood, C. R., Dinwiddie, G., Bailey, V., Carta, J. J., Dorsey, D., Kohler, F. W., . . . Schulte, D. (1987). Field replication of classwide peer tutoring. *Journal of Applied Behavior Analysis, 20,* 151–160.

*Greenwood, C. R., Dinwiddie, G., Terry, B., Wade, L., Stanley, S. O., Thibadeau, S., & Delquadri, J. C. (1984). Teacher-versus peer-mediated instruction: An ecobehavioral analysis of achievement outcomes. *Journal of Applied Behavior Analysis, 17,* 521–538.

Greenwood, C. R., Terry, B., Arreaga-Mayer, C., & Finney, R. (1992). The classwide peer tutoring program: Implementation factors moderating students' achievement. *Journal of Applied Behavior Analysis, 25,* 101–116.

*Hawkins, R. O., Musti-Rao, S., Hughes, C., Berry, L., & McGuire, S. (2009). Applying a randomized interdependent group contingency component to classwide peer tutoring for multiplication fact fluency. *Journal of Behavioral Education*, *18*, 300–318.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practices in special education. *Exceptional Children, 71*, 165–179.

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education, 21,* 266–272.

Horner, R. H., Swaminathan, H., Sugai, S., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35,* 269–290.

*Houghton, S., & Bain, A. (1993). Peer tutoring with ESL and below-average readers. *Journal of Behavioral Education, 3,* 125–142.

*Hughes, T. A., & Fredrick, L. D. (2006). Teaching vocabulary with students with learning disabilities using classwide PT and constant time delay. *Journal of Behavioral Education, 15,* 1–23.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage Publications.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108–446, 118 Stat. 2647(2004): 34 CFR §§300.10 *et seq.* (1975) (amending 20 U.S.C. §§ 1400 *et seq.*)

Jun, S. W., Ramirez, G., & Cumming, A. (2010). Tutoring adolescents in literacy: A meta-analysis. *Journal of Education, 45,* 219–238.

*Kamps, D. M., Berbetta, P. M., Leonard, B. R., & Delquadri, J. (1994). Classwide peer tutoring: An integration strategy to improve reading skills and promote peer interactions among students with autism and general education peers. *Journal of Applied Behavioral Analysis, 27,* 49–61.

*Kamps, D. M., Greenwood, C., Arreaga-Mayer, C., Veerkamp, M. B., Utley, C., Tapia, Y., Bowman-Perrott, L., & Bannister, H. (2008). The efficacy of ClassWide Peer Tutoring in middle schools. *Education and Treatment of Children, 31,* 119–152.

*Kourea, L., Cartledge, G., & Musti-Rao, S. (2007). Improving the reading skills of urban elementary students through total class peer tutoring. *Remedial and Special Education, 28,* 95–107.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation.* Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

*Lo, Y., & Cartledge, G. (2004). Total class peer tutoring and interdependent group oriented contingency: Improving the academic and task related behaviors of fourth-grade urban students. *Education and Treatment of Children, 27,* 235–262.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 3,* 339–359.

*Mackiewicz, S. M., Wood, C. L., Cooke, N. L., & Mazzotti, V. L. (2011). Effects of peer tutoring with audio prompting on vocabulary acquisition for struggling readings. *Remedial and Special Education, 32,* 345–354.

*Maheady, L., & Harper, G. F. (1987). A classwide peer tutoring program to improve the spelling performance of low-income, third-, and fourth-grade students. *Education and Treatment of Children, 10,* 120–133.

*Maheady, L., Harper, G. F., & Sacca, K. (1988). A classwide peer tutoring system in a secondary, resource room program for the mildly handicapped. *Journal of Research and Development in Education, 21,* 76–83.

*Maheady, L., Sacca, M. K., & Harper, G. F. (1987). Classwide student tutoring teams: The effects of peer mediated instruction on the academic performance of secondary mainstreamed students. *The Journal of Special Education, 21,* 107–121.

Mastropieri, M. A., Spencer, V., Scruggs, T. E., & Talbott, E. (2001). Students with disabilities as tutors: An updated research synthesis. *Advances in Learning and Behavioral Disabilities, 15,* 247–279.

*Mayfield, K. H., & Vollmer, T. R. (2007). Teaching math skills to at-risk students using home-based peer tutoring. *Journal of Applied Behavior Analysis, 40,* 223–237.

Mitchem, K. J., Young, K. R., West, R. P., & Benyo, J. (2001). CWPASM: A classwide peer assisted self-management program for general education classrooms. *Education and Treatment of Children, 24,* 111–140.

National Association of School Psychologists. (2010). *Model for comprehensive and integrated school psychological services.* Retrieved from http://www.nasponline.org/standards/2010standards/2_PracticeModel.pdf

*Neddenriep, C. E., Skinner, C. H., Wallace, M. A., & McCallum, E. (2009). Classwide peer tutoring: Two experiments investigating the generalized relationship between increased oral reading fluency and reading comprehension. *Journal of Applied School Psychology, 25,* 244–269.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.

*Oddo, M., Barnett, D. W., Hawkins, R. O., & Musti-Rao, S. (2010). Reciprocal peer tutoring and repeated reading: Increasing practicality using student groups. *Psychology in the Schools, 47,* 842–858.

Okilwa, N.S.A., & Shelby, L. (2010). The effects of peer tutoring on academic performance of students with disabilities in grades 6 through 12: A synthesis of the literature. *Remedial and Special Education, 31,* 450–463.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40,* 194–204.

Parker, R., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75,* 135–150.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining non-overlap and trend for single-case research: Tau-U. *Behavior Therapy, 35,* 202–322.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science, 3*(4), 1–6.

*Pigott, H. E., Fantuzzo, J. W., & Clement, P. W. (1986). The effects of reciprocal peer tutoring and group contingencies on the academic performance of elementary school children. *Journal of Applied Behavior Analysis, 19,* 93–98.

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. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 232–243). New York, NY: Russell Sage Foundation.

*Salend, S. J., & Nowak, M. R. (1988). Effects of peer-previewing on LD students' oral reading skills. *Learning Disability Quarterly, 11,* 47–53.

*Schloss, P. J., Kobza, S. A., & Alper, S. (1997). The use of peer tutoring for the acquisition of functional math skills among students with moderate retardation. *Education and Treatment of Children, 20,* 189–208.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8*(2), 24–33.

*Sutherland, K. S., & Snyder, A. (2007). Effects of reciprocal peer tutoring and self-graphing on reading fluency and classroom behavior of middle school students with emotional or behavioral disorders. *Journal of Emotional and Behavioral Disorders, 15,* 103–118.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44,* 423–432.

Utley, C. A., & Mortweet, S. L., (1997). Peer-mediated instruction and interventions. *Focus on Exceptional Children, 29*(5), 1–23.

Vannest, K. J., Parker, R. I., & Gonan, O. (2011). *Single-case research: Web-based calculators for SCR analysis, Version 1.0* (Web-based application). College Station, TX: Texas A&M University. Available from www.singlecaseresearch.org

*Veerkamp, M. B., Kamps, D. M., & Cooper, L. (2007). The effects of classwide peer tutoring on the reading achievement of urban middle school students. *Education and Treatment of Children, 30*(2), 21–51.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching: A multimedia training package.* Columbus, OH: Merrill.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Date Received: February 01, 2012
Date Accepted: December 17, 2012
Action Editor: Cynthia Anderson ■

Lisa Bowman-Perrott, PhD, is an assistant professor in the Department of Educational Psychology, Special Education, at Texas A&M University. Her primary research interests are academic and behavioral interventions for students with or at risk for emotional and behavioral disorders, including peer tutoring.

Heather S. Davis is a doctoral student in the Special Education program at Texas A&M University. Her research interests are identifying and providing evidence-based behavioral interventions to children exhibiting challenging behavior and training teachers and parents on implementing evidence-based behavioral interventions in school and home settings.

Kimber J. Vannest, PhD, is an associate professor in the Department of Educational Psychology, Special Education, at Texas A&M University. Her research interests are in determining effective interventions for children and youth with or at risk for emotional and behavioral disorders, including teacher behaviors and measurement.

Lauren Williams, MEd, is a special education teacher for students with communication, academic, and social learning deficits. Her primary interest is in working with students with emotional and behavioral disorders; she is working toward certification as a board-certified behavior analyst.

Charles R. Greenwood, PhD, is director of the Juniper Gardens Children's Project (JGCP), senior scientist in the Schiefelbusch Institute for Life Span Studies, and a professor in the Department of Applied Behavioral Science and the Department of Special Education at the University of Kansas. He is the developer of ClassWide Peer Tutoring, a class-wide instructional approach for teaching basic academic skills.

Richard Parker, PhD, is recently retired from Texas A&M University Educational Psychology Department. His continued research interest is single-case research methodology.