ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

Effects of Explicit Vocabulary Interventions for Preschoolers: An Exploratory Application of the Percent of Goal Obtained Effect Size Metric

Ella Patrona,[a] John Ferron,[a] Arnold Olszewski,[b] Elizabeth Kelley,[c] and Howard Goldstein[a]

[a] University of South Florida, Tampa [b] Miami University, Oxford, OH [c] University of Missouri, Columbia

ARTICLE INFO

Article History:
Received April 16, 2022
Revision received July 4, 2022
Accepted August 31, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Megan York Roberts

https://doi.org/10.1044/2022_JSLHR-22-00217

ABSTRACT

Purpose: Systematic reviews of literature are routinely conducted to identify practices that are effective in addressing educational and clinical problems. One complication, however, is how best to combine data from both group experimental design (GED) studies and single-case experimental design (SCED) studies. Percent of Goal Obtained (PoGO) has been developed as a metric to express the size of the effect relative to the distance to a goal, which could have broad applicability. This study sought to augment this descriptive index with estimates of standard errors, which are needed to use PoGO as an effect size metric in meta-analyses of SCED and GED studies. This study investigated the application of PoGO and standard errors to both SCED and GED studies examining a common intervention approach used with a single population.
Method: Sixteen articles investigating explicit vocabulary instruction applied to pre-K and kindergarten students were identified. PoGO and standard errors were calculated for variations of explicit vocabulary interventions. Evaluated interventions included six studies using exclusively an SCED, nine studies using a GED, and one that used both.
Results: PoGO was calculated for each treatment condition when applicable (i.e., alternating treatments designs). Standard errors and confidence interval limits also were calculated. PoGO effect size values ranged from 14.4% to 93.6%. PoGO for single-case experiments was 49.2% with a standard error of 7.26, and for group experiments, it was 30.8% with a standard error of 3.71.
Conclusion: Despite variation in the percentage of goal obtained across studies, the high degree of overlap in PoGO and standard errors between single-case and group experiments provides an indication that systematic reviews can apply this effect size metric to combine information obtained across experimental designs.

Educators, clinicians, and other practitioners are expected to provide efficacious and efficient services to their students and clients. Rather than relying on the findings of a single study to identify best practices, they increasingly rely on systematic reviews of the empirical studies that evaluate the effects of interventions that are designed to address educational and clinical problems of interest. Indeed, systematic reviews have become increasingly more important as a source of information for clinicians to identify empirically supported interventions (Schlosser & Sigafoos, 2008, 2009). One complication in identifying empirically supported interventions is the high likelihood that the review will include both group experimental design (GED) studies and single-case experimental design (SCED) studies. For example, Sanders et al. (2019) conducted a meta-analysis of self-regulated strategy development reading interventions to improve the reading comprehension of students with disabilities that included four GED studies and five SCEDs. Similarly, Wood et al. (2018) meta-analyzed

Correspondence to Howard Goldstein: hgoldstein@usf.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

22 studies, including both GEDs and SCEDs, to examine the effects of text-to-speech tools on the reading comprehension of students with reading disabilities.

These systematic reviews often use effect size metrics to estimate the size of treatment effects and aggregate them to identify evidence-based practices (Shadish et al., 2015). When both SCED and GED studies are to be included in the systematic review, two problems can arise, both related to the metric of the effect sizes used. First, the effect sizes commonly used for SCED studies are not comparable (i.e., not on the same scale) as the effect sizes commonly used with GED studies, which makes comparisons or aggregation across SCED and GED studies difficult. Second, for both SCED and GED studies, the metrics commonly used to index the size of the effect can be challenging to interpret and are not well aligned with the practical or clinical significance of the effect.

Because of the poor alignment in effect size estimates used for GED versus SCED studies, some researchers conduct systematic reviews on one type of design to the exclusion of other designs. However, reviews drawn only from a subset of the available evidence can be misleading. For example, Goldstein et al. (2014) conducted a review of the quality and effects of SCED and GED studies of social skills interventions for young children with autism. They argued that reviews of GED studies only wrongly concluded that no efficacious interventions were available. Thus, investigators continue to search for approaches that allow researchers to incorporate available evidence from both GED and SCED studies into systematic reviews.

Regarding the issue of comparability, most SCED effect sizes are based on variability within an individual, whereas GED effect sizes are based on variability among individuals. More specifically, SCED studies commonly use nonoverlap indices, such as Nonoverlap of All Pairs (Parker & Vannest, 2009) and Tau U (Parker et al., 2011), or mean differences (or trend adjusted mean differences) that have been standardized by the variation within the individual (Busk & Serlin, 1992). In contrast, GED studies typically use standardized mean differences, where the standard deviation is based on the between-persons variation, pooled across treatment and control groups. Because the variance between persons is larger than the variance within a person, standardized mean differences from GED studies are systematically smaller than the within-person standardized mean differences from SCED studies (Shadish et al., 2014; Van den Noortgate & Onghena, 2008). This difference in scale makes it difficult to aggregate or compare effect sizes from GED and SCED studies.

One approach for dealing with the lack of comparability is to meta-analyze the SCED studies separately from the GED studies. For example, in a meta-analysis of social and communication intervention practices for children with autism spectrum disorder, Bejarano-Martin et al. (2020) kept the two types of study separate, as did Jamieson et al. (2014) in their meta-analysis of cognitive prosthetic technology for individuals with memory impairments. When the outcomes in the GED studies are unlike the outcomes in the SCED studies, meta-analysts would often want to obtain separate effect size estimates for the different outcome types, and thus the separate analysis approach works well. However, when outcomes are similar across SCED and GED studies, the separate analysis approach limits the meta-analyst in a couple ways. First the average effect estimate from each type of study is based only on the data from one type of study and, thus, is less precise than an average effect estimate based on all the data. Second, moderator analyses are restricted to a subset of the data based on one design type, which limits the meta-analyst's power to identify factors that impact the effectiveness of the treatment.

A second approach for dealing with this problem is to use effect size measures for SCEDs, which, like GED studies, standardize based on an estimate of between-persons variability. For example, in a multiple-baseline study with three or more participants, it is possible to estimate the variability between cases, and then use this variability estimate to standardize the mean difference. These design-comparable effect sizes (D-CESs) have been developed (Hedges et al., 2012, 2013; Pustejovsky et al., 2014; Shadish et al., 2014; Swaminathan et al., 2014; Van den Noortgate & Onghena, 2008), and a web application is available to facilitate their computation (Pustejovsky et al., 2021).

However, there are limitations associated with the use of D-CES. One concern is that D-CESs require at least three individuals for estimation of the between-persons variability, and thus, effect evidence from high-quality SCED studies will be excluded from meta-analyses when those studies have only one or two participants (e.g., multiple-baseline designs across behaviors or settings, reversal designs, and alternating treatments designs). For example, Maggin et al. (2017) found in their meta-analysis of group contingency interventions on academically engaged and disruptive behaviors that the use of design-comparable effect sizes led to the exclusion of 13 of the 40 single-case studies that had met What Works Clearinghouse (2020) standards. Another concern with D-CESs is that they yield only an across-person average effect, as opposed to an effect estimate for each person, and thus, the variation in the response to intervention of the participants in a study cannot be explored as part of a moderator analysis. Finally, there are concerns with trying to interpret the clinical significance of a D-CES of a particular value.

The interpretation challenges associated with standardized mean differences are not unique to SCED studies. These concerns have also been voiced in the GED literature, where they have motivated the development of common language effect sizes (Brooks et al., 2014). The standardized mean difference tells us the size of the difference between the means in standard deviation units (e.g., the mean of the treatment group was 0.20 *SDs* higher than that of the control group). However, relating the size of the standardized mean difference to the clinical significance of the effect has proved elusive. This is unfortunate, because the goal of practitioners typically is to intervene in hopes of alleviating clinical concerns or improving academic achievement and behavior, so they are on par with typically developing peers. The goal may be the extinction of a problem behavior, or a 100% correct on a criterion measure (e.g., math or vocabulary test), or it could be a rate of social interaction that is derived normatively from observations of typical individuals. Depending on the behavioral phenomenon of interest, one would expect practitioners and researchers to be able to agree on a common goal.

Percent of Goal Obtained (PoGO) has been developed as a metric for SCED studies that expresses the size of the effect relative to the distance to the goal, and thus, a PoGO of 0 indicates no progress toward the goal, whereas a PoGO of 100 indicates the goal was obtained (Ferron et al., 2020). More formally, in the context of behavior acquisition, the following formula can be used to calculate PoGO:

$$\text{PoGO} = \frac{\beta - \alpha}{\gamma - \alpha} \times 100, \tag{1}$$

where γ is the goal level of behavior, α is the expected level of behavior without intervention, and β represents the actual level of behavior achieved during the intervention. This formula can also be adapted to be used in contexts where the intervention is designed for behavior reduction:

$$\text{PoGO} = \frac{\alpha - \beta}{\alpha - \gamma} \times 100. \tag{2}$$

By establishing a clinically relevant goal, this effect size can be more readily interpreted. In addition, the expected level of the behavior without intervention (α) and the level of behavior during intervention (β) can be comparably estimated from both SCED and GED studies, and thus, if a comparable goal is set for the SCED and GED studies, it provides an index that can be compared or aggregated across SCED and GED studies.

However, PoGO was developed as a descriptive index, and thus, no standard errors were derived and presented for PoGO. The lack of standard errors presents multiple problems for those that would like to use PoGO as an effect size metric in a meta-analysis of SCED and GED studies. First, standard errors are essential for estimating confidence intervals around effect estimates for individual studies (or individual participants in SCED studies). Second, standard errors for the study and case-specific PoGO estimates are needed to estimate the confidence interval for the meta-analytic mean effect estimate that is obtained by averaging effect sizes across studies. Third, the meta-analytic mean effect estimate can be estimated more precisely if the effect size estimates are weighted using the inverse of the error variance, which is based on the standard errors.

Purpose

The purpose of this study is to further the development of PoGO for systematic reviews. First, we aim to develop a method for estimating standard errors for PoGO. Second, we aim to explore the applicability of PoGO and the developed standard errors in systematic reviews by using them to synthesize a combination of SCED and GED studies that examined explicit vocabulary interventions for preschoolers.

Standard Errors and Confidence Intervals for PoGO

To derive standard errors for PoGO, we first distinguish between the quantity we are trying to estimate, PoGO, which is defined by α, β, and γ (as shown in Equations 1 and 2), and the estimate of that quantity, which relies on estimates of α (i.e., $\hat{\alpha}$) and β (i.e., $\hat{\beta}$). There are multiple ways of estimating α and β, and, as a consequence, multiple ways of estimating PoGO. For example, in a context where temporal stability is assumed, the mean of the baseline observations may be used to estimate α, but if temporal stability is only assumed for the last *n* observations of baseline, then the mean of the last *n* baseline observations may be used to estimate α, and if temporal stability is not assumed, but rather it is assumed that there is systematic change due to maturation, then a regression-based extension of the baseline trend may be used to estimate α (Ferron et al., 2020). How α is estimated will affect its standard error and also the standard error for PoGO. Similarly, how β is estimated will affect its standard error and the standard error for PoGO. Furthermore, the standard errors for PoGO will depend on assumptions about how the observations are distributed within a phase (e.g., normal or Poisson; independent or autocorrelated). Shadish and Sullivan (2011) found in a review of the characteristics of 113 SCED studies that there was considerable variability in the designs used, the outcomes examined, and the estimated autocorrelation, which suggests that the assumptions for analysis in one SCED context may differ from another.

To derive the standard errors, we chose to use the approximate formula for the standard error of a ratio of normal variables derived by Dunlap and Silver (1986). Doing so, involves a normality assumption about the estimate of the numerator of PoGO (i.e., $\hat{\beta} - \hat{\alpha}$) and the estimate of the denominator of PoGO (i.e., $\gamma - \hat{\alpha}$). These estimates are often based on means, mean differences, or time trend-adjusted mean differences (e.g., regression coefficients) and, thus, will tend toward normality as the number of observations increases. We substituted the PoGO numerator (i.e., $\hat{\beta} - \hat{\alpha}$) and denominator (i.e., $\gamma - \hat{\alpha}$) into Dunlap and Silver's approximate formula to obtain an approximate formula for the PoGO standard errors:

$$SE_{\text{PoGO}} = \frac{\sqrt{s^2_{\beta-\alpha} + \frac{\left(\hat{\beta}-\hat{\alpha}\right)^2}{(\gamma-\hat{\alpha})^2}\, s^2_{\gamma-\alpha}}}{\gamma - \hat{\alpha}}, \tag{3}$$

where $\hat{\alpha}$ is the estimate of α, $\hat{\beta}$ is the estimate of β, $s^2_{\beta-\alpha}$ is the error variance in estimating $\beta - \alpha$, and $s^2_{\gamma-\alpha}$ is the error variance in estimating $\gamma - \alpha$.

In some SCED contexts, researchers need to control for baseline trend (e.g., Parker et al., 2006), and in others, they need to account for autocorrelation (for discussions of the prevalence of autocorrelation in SCEDs see Matyas & Greenwood, 1996; Shadish & Sullivan, 2011). In these relatively complex contexts, Equation 3 could be used by first estimating the quantities used in PoGO (i.e., $\beta - \alpha$, and α) using a regression model that assumed an autoregressive error structure (e.g., generalized least squares regression, Maggin et al., 2011, or Bayesian methods, Swaminathan et al., 2014) and that included separate time trends for baseline and treatment phases. By centering the time variable in the regression (see, e.g., Huitema & McKean, 2000), regression coefficients could be obtained that would correspond to $\beta - \alpha$ at a particular focal time (e.g., the difference between the projected baseline trend line and the treatment phase trend line at the end of treatment) and α (e.g., the projected baseline value at that same focal time). The estimates of these regression coefficients along with the error variances for these coefficients could then be substituted into Equation 3 to estimate the standard error for PoGO.

In some SCED contexts, the estimates for the quantities in PoGO and its standard error can be made more simply. We may be able to assume we are dealing with a behavior in which there would be stable responding for baseline observations, and thus, the expected level of the behavior without intervention (α) could be estimated as the mean of the baseline observations. Similarly, we may be able to assume that the treatment observations would either be stable or stable at the end of the treatment phase, and thus, the level of behavior resulting from intervention (β) could be estimated as the mean of the treatment phase observations or the mean of the last *n* treatment observations. Finally, we may be examining an outcome where we have enough spacing between the observations to assume that autocorrelation would be negligible so that the observations could be assumed to be distributed independently. Under these assumptions, the error variance of the numerator of PoGO ($s^2_{\beta-\alpha}$) would be:

$$s^2_{\beta-\alpha} = \left(\frac{s^2_A}{n_A} + \frac{s^2_B}{n_B}\right), \tag{4}$$

where s^2_A is the variance of the participant's baseline observations, s^2_B is the variance of the participant's intervention observations that were used to compute the estimate of β, n_A is the number of baseline observations, and n_B is the number of intervention observations used to compute the estimate of β. Because γ is a known quantity, the estimated error variance in the denominator of PoGO under these simplifying assumptions is the error variance in the mean of the participant's baseline observations.

$$s^2_{\gamma-\alpha} = \frac{s^2_A}{n_A}. \tag{5}$$

By substituting these more specific estimators of the error variance from Equations 4 and 5 into Equation 3, we obtain a more specific formula for PoGO standard errors.

$$SE_{\text{PoGO}} = \frac{\sqrt{\left(\frac{s^2_A}{n_A} + \frac{s^2_B}{n_B}\right) + \frac{\left(\hat{\beta}-\hat{\alpha}\right)^2}{(\gamma-\hat{\alpha})^2}\frac{s^2_A}{n_A}}}{\gamma - \hat{\alpha}}. \tag{6}$$

Dunlap and Silver (1986) showed through simulations that their approximation to the standard error of a ratio of normal variables was accurate when the variables were uncorrelated and when the denominator of the ratio was relatively large compared with its error variance (i.e., at least 6 times as large). We found in our study to be presented next that the value of $\frac{\gamma-\alpha}{s^2_{\gamma-\alpha}}$ was at least six for 105 of the 106 PoGO standard error estimates (the one value less than six was 4.6).

The standard errors from either Equation 3 or 6 can be used to calculate confidence intervals, with the following formula:

$$\text{PoGO} \pm tcrit(SE), \tag{7}$$

where *tcrit* is substituted with the corresponding value from a *t* distribution table, based on the degrees of freedom in the study and the desired alpha level.

Study of the Applicability of PoGO and Its Standard Errors

The purpose of the applicability study is to explore the application of PoGO and its standard error to an array of studies that sought to teach vocabulary to preschoolers. In addition to SCED and cluster randomized design studies for which we had access to participants' data, we identified a number of published studies that also sought to teach vocabulary explicitly to preschoolers and kindergartners. This data set allowed us to investigate whether estimates of average effect sizes, standard errors, and confidence intervals were producing similar results. We also sought to investigate the process of using these data to produce an overall assessment of the efficacy of vocabulary interventions for preschoolers, with a focus on those who were at risk or demonstrating language delays.

Method

Article Selection Criteria

The primary goal of this study was to demonstrate the applicability of PoGO as a method of comparing effect sizes across GED and SCED studies. Thus, rather than an exhaustive literature search to identify all studies of early childhood vocabulary interventions, we selected a subset of articles published since 2000 that met the following criteria: (a) included preschool or kindergarten-aged children, (b) intervention that included explicit instruction on vocabulary words, (c) outcome measures assessed knowledge of the taught vocabulary words (i.e., studies that included only broad vocabulary measures, such as the PPVT, were excluded), and (d) reported or had available sufficient data to calculate PoGO. Both GED and SCED studies were included.

The authors initially identified 26 articles. Six articles were eliminated because there was not sufficient data reported to calculate PoGO. These articles were most often missing pretest data. Four articles were eliminated because outcome measures were not specific to taught words. It should be noted that several of the included articles were published by the authors of this article; therefore, we had access to raw data to be used for PoGO calculations. The remaining 16 articles included 17 studies: 10 GED and seven SCED (i.e., Kelley et al., 2015, included both an SCED and a GED).

Study Characteristics

We have summarized information about the 17 included studies in Table 1. Four categories of information were extracted to capture information relevant to the comparison of the studies. The categories included (a) the size and characteristics of the sample investigated, (b) the experimental design used, (c) the size and characteristics of the vocabulary to be taught, and (d) the types of measures of vocabulary learning used. Nine of the included studies (five SCED, four GED) reported results from iterations of the same intervention, *Story Friends*. The remaining eight studies reported results from other vocabulary interventions.

Data Analysis

Data were extracted from the results section of each study (or from raw data researchers could access) to calculate PoGO. Of the seven included SCED studies, four used a repeated acquisition design (RAD), two used a RAD with an adapted alternating treatment design (AATD), and one used a multiple-baseline design across participants. The four RAD studies contained observation-level data for each child (e.g., pre- and posttest vocabulary scores for each book in a nine-book series). For these studies, a PoGO estimate was obtained for each participant using the pretest mean as the estimate of α and the posttest mean as the estimate of β. The goal level, represented by γ, was set to the total possible number of points that the child could earn in the vocabulary knowledge measure. For the two RAD and AATD design studies (Dennis & Whalon, 2021; Seven et al., 2020), the PoGO estimate was calculated in the same way for each treatment condition. For example, a separate POGO value was calculated for the automated application treatment and the teacher-delivered treatment in Dennis and Whalon (2021). For the multiple-baseline design (Bobzien et al., 2015), a PoGO value was estimated for each book for each child, using the mean of the baseline observations for the book as the estimate of α and the mean of the last three treatment observations within the book as the estimate β. The standard errors for each PoGO value were then estimated using Equation 6, which was programmed into an Excel spreadsheet. Our rationale for using Equation 6 was that we were using means as estimates of α and β given the observed stability in responding, and because the SCEDs were mostly RADs, which have enough spacing between the observations to limit concerns with autocorrelation.

We then obtained a meta-analytic mean PoGO estimate for each SCED. For RADs, the case-specific PoGO values were aggregated into a summary value for the study using a mixed linear model. More specifically, we used a meta-analytic model:

$$\widehat{PoGO}_{ij} = PoGO_j + u_{ij} + e_{ij}, \tag{8}$$

where $\widehat{PoGO}_{ij}$ is the estimated PoGO for Participant i in Study j, $PoGO_j$ is the true PoGO for Study j, u_{ij} is the deviation of the true PoGO for Participant i in Study *j* from the true PoGO for Study j, and e_{ij} is the residual that accounts for the error in the estimate introduced by the sampling of the observations. Both u_{ij} and e_{ij} are assumed normally distributed. The model was estimated using SAS 9.4, and program code was adapted from Konstantopoulos (2011); see the first set of programming code in the Appendix for more information.

For the multiple-baseline study, we used a similar meta-analytic model where the estimated PoGO for

Table 1. Summaries of the articles selected for analysis.

Authors	Participants	Design and intervention length	Vocabulary selection	Vocabulary knowledge measure
Spencer et al. (2012)	*N* = 9; inclusion criteria: PPVT-4 and/or CELF-P2 of 78–92; preschool; all English-speaking children from low-income families	Repeated acquisition single-case experimental design; approximately 9 weeks of intervention	18 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question.
Bobzien et al. (2015)	*N* = 4; inclusion criteria: congenital hearing loss, PLS-5 or PPVT-4 scores below age expectations, narrative skills below age expectations; all English-speaking children	Multiple-baseline design replicated across children; 8–30 intervention sessions	30 words Tiers 1 and 2 vocabulary, short phrases	Expressive production task: Children responded to open-ended questions to elicit use of target words.
Kelley et al. (2015)	*N* = 18; inclusion criteria: PPVT-4 standard score of 90 or lower; preschool; all English-speaking children; included children from low-income families	Randomized control trial with embedded repeated acquisition single-case experimental design; BAU comparison condition; 14 weeks	18 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question.
Greenwood et al. (2016)	*N* = 9; inclusion criteria: below cut point on IGDI Picture Naming; preschool; included English language learners and children from low-income families	Repeated acquisition single-case experimental design (replication of Spencer et al., 2012); approximately 9 weeks	18 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question.
Peters-Sanders et al. (2020)	*N* = 17; inclusion criteria: PPVT-4 or CELF-P2 of 70–130; preschool; all English-speaking children; included children from low-income families	Repeated acquisition single-case experimental design; 9 weeks	36 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definition question.
Seven et al. (2020)	*N* = 23; inclusion criteria: PPVT-4 and/or CELF-P2 0–1.5 *SD*s below mean; preschool; all English-speaking children; included children from low-income families	Repeated acquisition and adapted alternating treatment single-case experimental design; 11 weeks	32 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definition question.
Dennis & Whalon (2021)	*N* = 6; inclusion criteria: teacher identification and PLS-5 total language score below 35th percentile; preschool; all English-speaking children	Repeated acquisition single-case experimental design; 8 weeks	48 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definition question; follow-up questions for example and demonstration.
Goldstein et al. (2016)	*N* = 163; inclusion criteria: PPVT-4 of 78–85 (extended to 71–96 in some classrooms); preschool; included English language learners and children from low-income families	Cluster randomized design; comparison received *Story Friends* without embedded lessons; 26 weeks	36 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question.
Kelley et al. (2020)	*N* = 84; inclusion criteria: PPVT-4 and/or CELF-P2 of 70–92; preschool; all English-speaking children; Included children from low-income families	Cluster randomized design; comparison received *Story Friends* without embedded lessons; 13 weeks	36 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definition question.
Madsen et al. (2022)	*N* = 84; inclusion criteria: PPVT-4 and/or CELF-P2 of 70–92; preschool; all English-speaking children; included children from low-income families	Cluster randomized design; comparison received *Story Friends* without embedded lessons; 13 weeks	36 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definition question.

(*table continues*)

Table 1. *(Continued).*

Authors	Participants	Design and intervention length	Vocabulary selection	Vocabulary knowledge measure
Justice et al. (2005)	*N* = 57; inclusion criteria: PALS-K score below median for child's school; Kindergarten; all English-speaking children; included children from low-income families	Pre/post randomized experimental design; BAU comparison condition; 10 weeks	60 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question; follow-up prompt to provide synonym for target word.
Coyne et al. (2010)	*N* = 124; kindergarten; included English language learners	Pre/post quasi-experimental design; BAU comparison condition; 18 weeks	54 words Tier 2 vocabulary	Expressive definition task: Children respond to open-ended definitional question; follow-up question with target word in neutral context.
Neuman & Dwyer (2011)	*N* = 178; voluntary; preschool; all English-speaking children; included children from low-income families	Pre/post quasi-experimental design; BAU comparison condition; 16 weeks	130 words Tier 2 vocabulary and "partially familiar" words	Expressive naming test: Children shown picture cards and name each picture.
Dickinson et al. (2019); Study 2	*N* = 84; voluntary; preschool; included English language learners and children from low-income families	Within-subject design; BAU comparison condition; approximately 8 weeks	64 words Selected based on previous work of Biemiller (2010) and Dickinson & Tabors (2001)	Receptive task: Children select picture of target word; choices were correct referent, thematically related foil, and conceptually related foil.
Zucker et al. (2019)	*N* = 1,193; voluntary; preschool and kindergarten (reported and analyzed separately); included English language learners and children from low-income families	Posttest only experimental design; BAU comparison condition; 26 weeks	227 words Tier 2 vocabulary	Receptive task: Children select picture of target word; choices were correct referent and two foils.
Zucker et al. (2021)	*N* = 167; voluntary; preschool; all dual language learners (English and Spanish; included children from low-income families	Randomized control trial; BAU comparison condition; 11 weeks	60 words Tier 2 vocabulary Intervention in Spanish	Receptive task: Children select picture of target word; choices were correct referent and two foils.

Note. PPVT-4 = Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007)); CELF-P2 = Clinical Evaluation of Language Fundamentals Preschool–Second Edition (Wiig et al., 2004); Tier 2 vocabulary = from Beck et al. (2002); PLS-5 = Preschool Language Scales–Fifth Edition (Zimmerman et al., 2011); IGDI = Individual Growth and Development Indicators (Bradfield et al., 2014); PALS-K = Phonological Awareness Literacy Screening–Kindergarten (Invernizzi et al., 2000); BAU = business as usual.

each book was aggregated to get a PoGO value for each case.

$$\widehat{PoGO}_{bi} = \mathrm{PoGO}_i + v_{bi} + e_{bi}, \qquad (9)$$

where $\widehat{PoGO}_{bi}$ is the estimated PoGO value for the Book b for Participant i, $PoGO_i$ is the true PoGO value for Participant i, v_{bi} is the deviation of the true PoGO for Book b for Participant i from the true PoGO for Participant i, and e_{bi} is the residual that accounts for the error in the estimate introduced by the sampling of the observations. The case-specific PoGO estimates were then aggregated as before to obtain a PoGO estimate for the study.

For all 10 group experimental studies, a PoGO value for the study was calculated using the posttest treatment and control group means and standard deviations. The control group's posttest mean indicated the expected level without intervention (α), and the treatment group's posttest mean indicated the level achieved during intervention (β; Ferron et al., 2020). Again, the goal level γ was set to the total possible number of points that the child could earn in the vocabulary knowledge measure. The standard errors for the GEDs were based on Equations 4 and 5, which assume random sampling. This approach likely underestimated the standard errors for the three GEDs based on cluster randomized controlled trials, a limitation that we will return to in our discussion.

After obtaining a PoGO estimate for each study, these values were then aggregated using similar meta-analytic models, one to obtain a meta-analytic mean PoGO value for the SCED studies, one for the GED studies, and one for all studies (see the second set of programming code in the Appendix, for an example). Finally, we used a mixed linear model to conduct a moderator analysis, in which we examined whether the size of the effect was moderated by the design type (see the third set of programming code in the Appendix).

Data coding and analysis were completed by the first and second authors. The third author independently checked data entry and analysis for reliability. This included returning to the original articles or data sets to ensure data were extracted and entered into spreadsheets appropriately and ensuring that formulas in Excel and SAS were entered correctly. A few minor discrepancies were noted and resolved in a meeting between the first and third authors, resulting in 100% agreement.

Results

Study Characteristics

Table 1 provides a summary of the included articles. The 16 articles included 17 studies with a variety of study designs, including GED (e.g., cluster-randomized, quasi-experimental designs; n = 10) and SCED (i.e., repeated acquisition single-case experimental designs: n = 4; repeated acquisition with adapted alternating treatments design: n = 2, multiple-baseline design across children: n = 1). POGO values were calculated for each study, and in the cases of Seven et al. (2020) and Dennis and Whalon (2021), POGO values were calculated for each treatment condition in the alternating treatments design. Thus, we had 17 studies that provided 19 PoGO values. Sample sizes ranged from four children in an SCED study to 1,193 children in a group design study. Many studies focused on children with language abilities below age expectations (n = 10), and many studies included participants from varying backgrounds, including English language learners (n = 6) and children from low-income households (n = 13). Many studies recruited participants at the classroom or district level and invited all children in those classrooms to participate. Other studies reported specific inclusionary criteria for participants (e.g., children with congenital hearing loss and weak expressive vocabulary; Bobzien et al., 2015). Studies examining the *Story Friends* intervention generally reported inclusionary criteria: Children were included based on their scores on the Peabody Picture Vocabulary Test–Fourth Edition (Dunn & Dunn, 2007), the Clinical Evaluation of Language Fundamentals Preschool–Second Edition (Wiig et al., 2004), or the Individual Growth and Development Indicators (Bradfield et al., 2014). Two other studies reported inclusion criteria (Dennis & Whalon, 2021; Justice et al., 2005) using scores on the Preschool Language Scales–Fifth Edition (PLS-5; Zimmerman et al., 2011) and the Phonological Awareness Literacy Screening–Kindergarten (PALS-K; Invernizzi et al., 2000), respectively.

The approach to selecting vocabulary targets was relatively consistent across all 17 studies. Most interventions referred to the work of Beck et al. (2002, 2013) and described vocabulary targets as "Tier 2" words. Tier 2 words, originally defined in Beck et al. (2002), refers to high-utility words that are not common to children's everyday interactions, but are likely to be presented in academic settings. Tier 2 words are hypothesized to be important for preparing children to learn to read. Bobzien et al. (2015) was a notable exception, as they included both Tier 1 and Tier 2 words. The number of words taught varied widely from 18 to 227 new words. Vocabulary knowledge was most often assessed through an expressive definition task in which children were asked to verbally define words and were awarded points based on complete, partial, or no knowledge of the target vocabulary word. Other studies assessed vocabulary knowledge using receptive tasks in which children were shown pictures of the correct referent and foils and asked to select the picture associated with the target word. Bobzien et al.

(2015) used an expressive production task in which children responded to open-ended questions designed to evoke target words.

Effect Sizes

Table 2 summarizes the PoGO effect size values obtained for each article. Recall that PoGO was calculated for each treatment condition, when applicable (i.e., the alternating treatments design in Seven et al. (2020). In addition to each PoGO value, the standard error and confidence interval limits also were calculated. Two studies (Dennis & Whalon, 2021; Seven et al., 2020) contrasted modifications in intervention delivery; thus, PoGO values for each treatment are reported. An average PoGO value also was calculated across all single-subject designs, all group designs, and all studies.

For each SCED and GED study, the PoGO value, standard error, and lower and upper confidence interval limits are provided in Table 2. Figure 1 presents a graphic summary of these data including the PoGO values and confidence intervals for each study. The PoGO values ranged from 14% to 93%. Confidence intervals were narrower, indicating more precise estimates of PoGO, for studies like Zucker et al.'s (2019) and Seven et al.'s (2020), whereas somewhat wider intervals were found for Zucker et al.'s (2021) and Dennis and Whalon's (2021). This figure also includes the average PoGO value for each type of study (SCED or GED), as well as an average across all studies. PoGO averaged across nine single-case treatments was 49.2% with an *SE* of 7.26. PoGO averaged across 10 group design studies was 30.8% with an *SE* of 3.71. PoGO averaged across all 18 studies was 40.0% with an *SE* of 4.48.

To more formally examine the difference in average PoGO values between the SCED and GED studies, we conducted a moderator analysis, where design type was entered as a predictor of the effect sizes in the mixed linear model. The difference was statistically significant, $\beta = -18.06$, $t(15) = -2.24$, $p = .041$, which suggests that, on average, a larger percent of the goal was obtained in the SCEDs.

Discussion

First, we reported results of a method for estimating standard errors for PoGO. Deriving the formulas for estimating the standard errors was an important step in the development of PoGO for the synthesis of literature because these standard errors are essential in estimating the confidence intervals for the PoGO effect size, in weighting the PoGO values to determine a meta-analytic mean, and in testing whether the effect is moderated by other factors. We derived both a more general equation for the PoGO standard errors (Equation 3) and a more specific equation (Equation 6). The more specific equation is limited to contexts with temporally stable within-phase

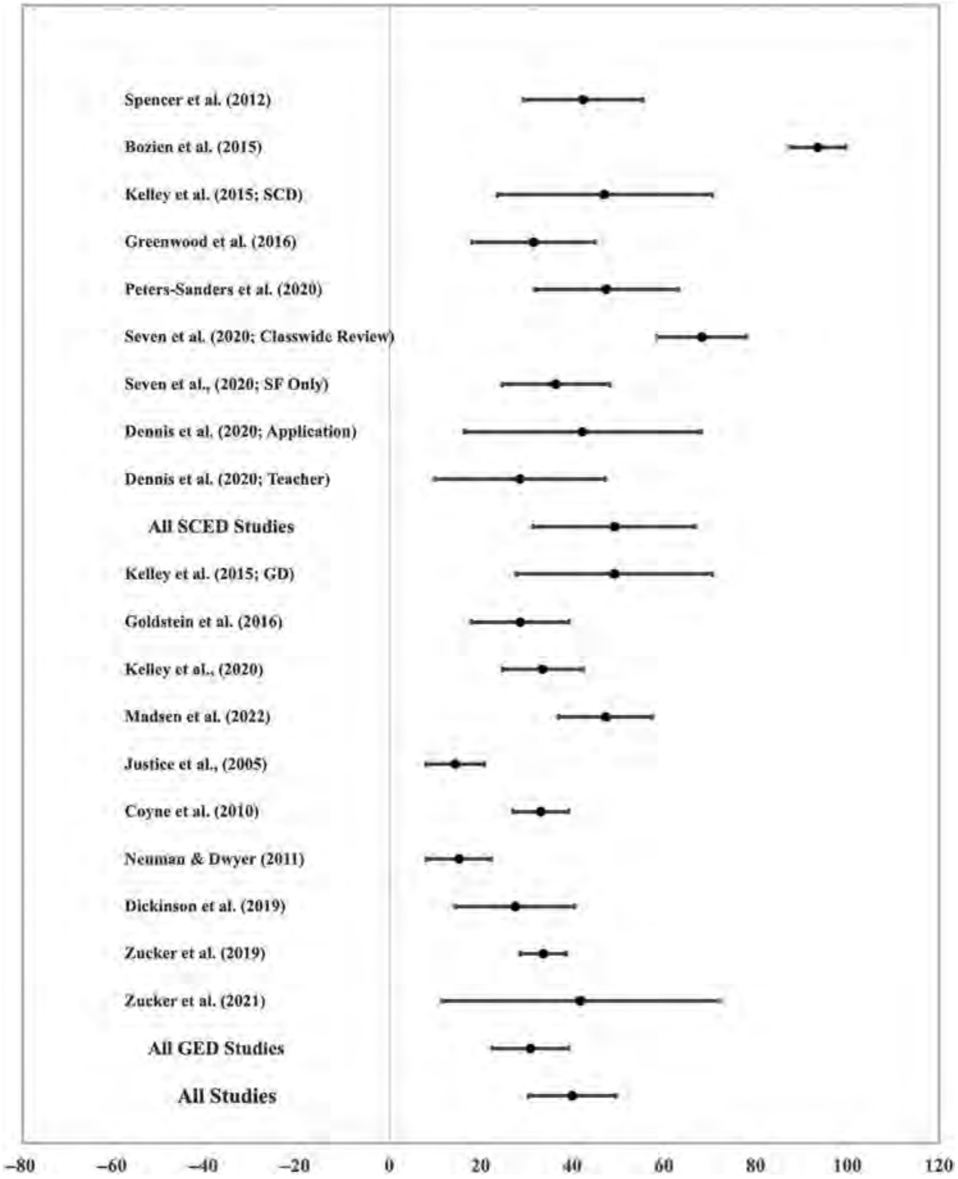
Table 2. PoGO results with standard errors and confidence interval limits.

Authors	PoGO	Standard error	Lower limit	Upper limit
Spencer et al. (2012)	42.3	5.61	29.4	55.3
Bobzien et al. (2015)	93.6	1.91	87.5	99.6
Kelley et al. (2015)[a]	47.0	10.2	23.6	70.4
Greenwood et al. (2016)	31.5	5.86	18.0	45.0
Peters-Sanders et al. (2020)	47.4	7.38	31.7	63.0
Seven et al. (2020): Classwide review strategies	68.2	4.74	58.3	78
Seven et al. (2020): *Story Friends* only	36.3	5.66	24.6	48.1
Dennis & Whalon (2021): Application	42.2	10.1	16.4	68.1
Dennis & Whalon (2021): Teacher	28.5	7.25	9.91	47.2
All single case experimental designs	49.2	7.26	31.4	66.9
Kelley et al. (2015)[a]	49.2	10.1	27.8	70.5
Goldstein et al. (2016)	28.6	5.46	17.9	39.3
Kelley et al. (2020)	33.5	4.40	24.8	42.3
Madsen et al. (2022)	47.2	5.14	36.9	57.4
Justice et al. (2005)	14.4	3.21	7.93	20.8
Coyne et al. (2010)	33.1	3.12	27.0	39.2
Neuman & Dwyer (2011)	15.2	3.59	8.13	22.2
Dickinson et al. (2019)	27.5	6.55	14.5	40.5
Zucker et al. (2019)	33.6	2.57	28.6	38.6
Zucker et al. (2021)	41.7	15.4	11.6	71.9
All group experimental designs	30.8	3.71	22.4	39.2
All studies	40.0	4.48	30.5	49.5

Note. Recall that PoGO is a measure of Percent of Goal Obtained, typically ranging from 0% to 100%.

[a] Kelley et al. (2015) is included in both categories due to the randomized group design with embedded single-case experimental design.

Figure 1. Percent of Goal Obtained and confidence interval for each study. SCED & SCD = single-case experimental design; GED & GD = group experimental design; SF = *Story Friends*.



responding and independent observations. Both equations are based on approximations to the standard error of a ratio of normal variables derived by Dunlap and Silver (1986). It will be important in future research (e.g., using simulation methods) to index the quality of these approximations with various types of single-case data, which may not be strictly normal. Our exploration of the applicability of PoGO and its standard errors for SCEDs was limited to a context where we made assumptions of temporal stability and independence and utilized Equation 6. Extensions of this work to more complex scenarios involving trends and/or autocorrelation where generalized least squares or Bayesian methods would be used to estimate the quantities in PoGO and their error variances are needed. Similarly, the standard errors we used for the GED studies assumed simple random sampling. Future research needs to extend the estimation of the error variances for the numerator and denominator of PoGO so that the clustering of data in cluster randomized control studies can be taken into account. Another future step in the development of PoGO for research synthesis would be to capitalize on the relationship between PoGO and response ratios (see Ferron et al., 2020) to develop ways to convert the results of meta-analyses conducted using log response ratios on to the PoGO scale for interpretation and vice versa.

The PoGO standard errors depend on the number of observations and the variability in the observations. For GED studies, the standard errors for PoGO will tend to be smaller when the sample size is larger and the population more homogeneous. For a single case from an SCED study, the PoGO standard error will tend to be smaller when there are more observations (i.e., a longer series length) and less variation within a phase, and also when there are more cases and less variation in the PoGO values across these cases. For example, Zucker et al. (2019) had a much larger sample size (*N* = 1,193) than Zucker et al. (2021; *N* = 167), resulting in a smaller standard error. A similar pattern can be observed in the SCED studies; Seven et al. (*N* = 23; 2020) had both a larger sample size and a smaller standard error than Dennis and Whalon (*N* = 6; 2021). In addition, in the multiple-baseline study (Bobzien et al., 2015), the number of observations per case was greater than in the other SCEDs plus the PoGO values were almost the same for each case, leading to a relatively small standard error for that study. Other study characteristics, such as the population of participants and the components of the intervention, may introduce variation in study outcomes and, thus, increase the standard error. For example, Zucker et al. (2021) examined the effects of a shorter, condensed version of the intervention used in Zucker et al. (2019), and enrolled English language learners in bilingual classrooms; these factors may have contributed to variability in children's learning in intervention and resulted in the higher standard error.

Second, we explored the applicability of PoGO and the confidence intervals, derived from standard errors, to synthesize a combination of SCED and GED studies that examined explicit vocabulary interventions. As you can see in the forest plot in Figure 1, there is a great deal of overlap in PoGO and confidence intervals among studies. There is one study with a particularly high PoGO estimate (Bobzien et al., 2015), where the confidence interval is outside the average PoGO and associated confidence interval for SCED studies. The high value for this study may be attributed to intervention characteristics (i.e., teaching Tier 1 and Tier 2 words and common phrases, as opposed to just Tier 2 words) or the measurement approach (i.e., children were asked to produce target words, as opposed to provide definitions). As well, the Bobzien study was the only multiple-baseline study included in our analysis; it may be that this higher PoGO value is related to a study design that extended intervention until a mastery criterion was met. Across studies, POGO effect sizes were generally larger when treatments were more intense. For example, in Seven et al. (2020), POGO was higher when words were taught with both small-group intervention and classwide review. In comparison, the two studies with low PoGO estimates (Justice et al., 2005; Neuman & Dwyer, 2011) provided less intense explicit instruction on target vocabulary words than other studies, which may account for their smaller PoGO values. Looking across SCED and GED studies, we see considerable overlap in the confidence intervals, but the average PoGO for the SCED studies (49.2) was larger than the average value from the GED studies (30.8), as shown in the moderator analysis. Overall, it appears that PoGO is readily applicable to both SCED and GED and offers some transparency in interpreting results as per mastery of the vocabulary words being taught.

Although differential effects are difficult to detect with a small literature base, it may be possible to explore hypothesized differences in PoGO that may be attributed to contrasting interventions or outcome measures. For example, one might hypothesize that teaching more words concurrently may have deleterious effects once the number overly taxes children's memory. One might hypothesize that instruction that adds a home practice component results in more vocabulary learning, reflected in a higher PoGO value. It is unclear how PoGO values compare when a receptive vocabulary measure with a higher chance pretest is contrasted with an expressive naming vocabulary measure. PoGO is likely to have value for conducting meta-analytic reviews for a variety of behavioral phenomena. Few preschool children master all the challenging new vocabulary words introduced in these explicit teaching experiments. Other behavioral phenomenon might be maximized to a greater extent, for example, interventions to curtail aggressive behavior or interventions to teach math concepts. Exploring the applicability of PoGO as a transparent measure of effects applicable to both single-case and group design experiments awaits further research.

For PoGO to be calculated from either SCED or GED studies, three items need to be available: (a) the expected value of the outcome in the absence of intervention, (b) the obtained value of the outcome during intervention, and (c) the goal level for the behavior. In single-case designs that have baselines and in group designs that have a control group, the expected value of the outcome in the absence of intervention is readily available (e.g., the mean of the control group for group designs, or the mean baseline values for single-case designs). However, PoGO cannot be calculated when the design does not provide such information. For example, PoGO could not be calculated for an alternating treatment design that rapidly alternates between Treatment A and B, with no initial baseline phase and no inclusion of the baseline condition in the alternating sequence. Similarly, PoGO cannot be calculated without a control group or pretest data in GED studies, which is why a few studies initially identified for this analysis were excluded. Also needed to compute PoGO is the obtained value of the outcome during treatment, something that is readily available for intervention studies (e.g., the posttreatment mean in a group study or the intervention values from a single-case study).

Finally, the goal value is needed. However, goal values are not always explicitly reported. We recommend that researchers regularly report the goal of their intervention, because doing so aids in the interpretation of the effects within a particular study and facilitates the synthesis of effects across studies using PoGO. In syntheses of single-case and group design studies using PoGO, it is critical that the goal value is comparable across studies (e.g., the goal of both the single-case and group design studies is 100% mastery of vocabulary words). Methods for setting goals for PoGO include using the ideal outcome across cases and studies (e.g., 100% mastery, 0% problem behavior) or using a standard normative value derived from the behavior of typically developing peers (see Ferron et al., 2020, for further elaboration of these methods). It would not be appropriate for the goal value to vary from case to case or from study to study based on researcher's judgments about a reasonable goal for a particular case, because different judgments would introduce noncomparability in the effect size estimates.

In conclusion, PoGO has the potential to add more transparency to effect size measures. Interventions are typically undertaken to improve or optimize behavior or learning. Among individuals with substantial deficits, interventions may result in large effects on measures of standardized mean differences. However, that should not be interpreted as evidence that one has achieved changes that ameliorate deficits and that performance of a sample of individuals receiving treatment now presents as normal. PoGO has the potential to capture the extent of behavior change in a way that can be related more directly to such a standard or a goal. Moreover, the addition of standard error calculations advances our ability to aggregate meta-analytic findings across studies GED and SCED studies using a common metric. PoGO holds promise as a useful effect size measure that is worthy of further application and investigation.

Author Contributions

Ella Patrona: Conceptualization (Supporting), Data curation (Lead), Formal analysis (Equal), Investigation (Lead), Visualization (Equal), Writing – review & editing (Supporting). **John Ferron:** Conceptualization (Equal), Formal analysis (Lead), Project administration (Equal), Writing – original draft (Equal). **Arnold Olszewski:** Conceptualization (Supporting), Data curation (Supporting), Formal analysis (Supporting), Investigation (Supporting), Validation (Equal), Writing – review & editing (Equal). **Elizabeth Kelley:** Conceptualization (Supporting), Data curation (Supporting), Formal analysis (Supporting), Writing – review & editing (Equal). **Howard Goldstein:** Conceptualization (Equal), Funding acquisition (Lead), Project administration (Lead), Visualization (Equal), Writing – original draft (Equal).

Data Availability Statement

The data sets analyzed for this review are available from the corresponding author on reasonable request.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A200179 to the University of South Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). Guilford.

Bejarano-Martin, A., Canal-Bedia, R., Magan-Maganto, M., Fernandez-Alvarez, C., Loa-Jonsdottir, S., Saemundsen, E., Vincente, A., Café, C., Rasga, C., Garcia-Primo, P., & Posada, M. (2020). Efficacy of focused social and communication intervention practices for young children with autism spectrum disorder: A meta-analysis. *Early Childhood Research Quarterly, 51,* 430–445. https://doi.org/10.1016/j.ecresq.2020.01.004

Biemiller, A. (2010). *Words worth teaching: Close the vocabulary gap*. McGraw Hill.

Bobzien, J. L., Richels, C., Schwartz, K., Raver, S. A., Hester, P., & Morin, L. (2015). Using repeated reading and explicit instruction to teach vocabulary to preschoolers with hearing loss. *Infants and Young Children, 28*(3), 262–280. https://doi.org/10.1097/iyc.0000000000000039

Bradfield, T. A., Besner, A. C., Wackerle-Hollman, A. K., Albano, A. D., Rodriguez, M. C., & McConnell, S. R. (2014). Redefining individual growth and development indicators. *Assessment for Effective Intervention, 39*(4), 233–244. https://doi.org/10.1177/1534508413496837

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology, 99*(2), 332–340. https://doi.org/10.1037/a0034745

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Erlbaum.

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Jr., Ruby, M., Crevecoeur, Y. C., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness, 3*(2), 93–120. https://doi.org/10.1080/19345741003592410

Dennis, L. R., & Whalon, K. J. (2021). Effects of teacher-versus application-delivered instruction on the expressive vocabulary of at-risk preschool children. *Remedial and Special Education, 42*(4), 195–206. https://doi.org/10.1177/0741932519900991

Dickinson, D. K., Nesbitt, K. T., Collins, M. F., Hadley, E. B., Newman, K., Rivera, B. L., Ilgez, H., Nicolopoulou, A., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Teaching for breadth and depth of vocabulary knowledge: Learning from explicit and implicit instruction and the storybook texts. *Early Childhood Research Quarterly, 47,* 341–356. https://doi.org/10.1016/j.ecresq.2018.07.012

Dickinson, D. K. & Tabors, P. O. (Eds.). (2001). *Beginning literacy with language: Young children learning at home and school.* Brookes.

Dunlap, W. P., & Silver, N. C. (1986). Confidence intervals and standard errors for ratios of normal variables. *Behavior Research Methods, Instruments, & Computers, 18*(5), 469–471. https://doi.org/10.3758/BF03201412

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. NCS Pearson Assessments. https://doi.org/10.1037/t15144-000

Ferron, J., Goldstein, H., Olszewski, A., & Rohrer, L. (2020). Indexing effects in single-case experimental designs by estimating the percent of goal obtained. *Evidence-Based Communication Assessment and Intervention, 14*(1–2), 6–27. https://doi.org/10.1080/17489539.2020.1732024

Goldstein, H., Kelley, E., Greenwood, C., McCune, L., Carta, J., Atwater, J., Guerrero, G., McCarthy, T., Schneider, N., & Spencer, T. (2016). Embedded instruction improves vocabulary learning during automated storybook reading among high-risk preschoolers. *Journal of Speech, Language, and Hearing Research, 59*(3), 484–500. https://doi.org/10.1044/2015_JSLHR-L-15-0227

Goldstein, H., Lackey, K. C., & Schneider, N. J. (2014). A new framework for systematic reviews: Application to social skills interventions for preschoolers with autism. *Exceptional Children, 80,* 262–280. https://doi.org/10.1177/0014402914522423

Greenwood, C. R., Carta, J. J., Guerrero, G., Atwater, J., Kelley, E. S., Kong, N. Y., & Goldstein, H. (2016). Systematic replication of the effects of a supplementary, technology-assisted, storybook intervention for preschool children with weak vocabulary and comprehension skills. *The Elementary School Journal, 116*(4), 574–599. https://doi.org/10.1086/686223

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224–239. https://doi.org/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*(4), 324–341. https://doi.org/10.1002/jrsm.1086

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*(1), 38–58. https://doi.org/10.1177/00131640021970358

Invernizzi, M., Meier, J. D., Swank, L., & Juel, C. (2000). *Phonological awareness literacy screening: Kindergarten*. University of Virginia.

Jamieson, M., Cullen, B., McGee-Lennon, M., Brewster, S., & Evans, J. J. (2014). The efficacy of cognitive prosthetic technology for people with memory impairments: A systematic review and meta-analysis. *Neuropsychological Rehabilitation, 24*(3–4), 419–444. https://doi.org/10.1080/09602011.2013.825632

Justice, L., Meier, J., & Walpole, S. (2005). Learning new words from storybooks: An efficacy study with at-risk kindergarteners. *Language, Speech, and Hearing Services in Schools, 36*(1), 17–32. https://doi.org/10.1044/0161-1461(2005/003)

Kelley, E. S., Barker, R. M., Peters-Sanders, L., Madsen, K., Seven, Y., Soto, X., Olsen, W. L., Hull, K., & Goldstein, H. (2020). Feasible implementation strategies for improving vocabulary knowledge of high-risk preschoolers: Results from a cluster-randomized trial. *Journal of Speech, Language, and Hearing Research, 63*(12), 4000–4017. https://doi.org/10.1044/2020_JSLHR-20-00316

Kelley, E. S., Goldstein, H., Spencer, T., & Sherman, A. (2015). Effects of automated tier 2 storybook intervention on vocabulary and comprehension learning in preschool children with limited oral language skills. *Early Childhood Research Quarterly, 31,* 47–61. https://doi.org/10.1016/j.ecresq.2014.12.004

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Madsen, K. M., Peters-Sanders, L. A., Kelley, E. S., Barker, R. M., Seven, Y., Olsen, W. L., Soto-Boykin, X., & Goldstein, H. (2022). Optimizing vocabulary instruction for preschool children. *Journal of Early Intervention*. https://doi.org/10.1177/10538151221116596

Maggin, D. M., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education, 38*(6), 353–370. https://doi.org/10.1177/0741932517716900

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*(3), 301–321. https://doi.org/10.1016/j.jsp.2011.03.004

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Psychology Press.

Neuman, S. B., & Dwyer, J. (2011). Developing vocabulary and conceptual knowledge for low-income preschoolers. *Journal of Literacy Research, 43*(2), 103–129. https://doi.org/10.1177/1086296X11403089

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*(4), 418–444. https://doi.org/10.1037/h0084131

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. (2011). Combining non-overlap and trend for single case research: Tau-U. *Behavior Therapy, 42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Peters-Sanders, L. A., Kelley, E. S., Biel, H. C., Madsen, K., Soto, X., Seven, Y., Hull, K., & Goldstein, H. (2020). Moving forward four words at a time: Effects of a supplemental preschool vocabulary intervention. *Language, Speech, and Hearing Services in Schools, 51*(1), 165–175. https://doi.org/10.1044/2019_LSHSS-19-00029

Pustejovsky, J. E., Chen, M., & Hamilton, B. (2021). *scdhlm: A web-based calculator for between-case standardized mean differences* (Version 0.5.2) [Web application]. https://jepusto.shinyapps.io/scdhlm

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework.. *Journal of Educational and Behavioral Statistics, 39*(5), 368–393. https://doi.org/10.3102/1076998614547577

Sanders, S., Losinski, M., Ennis, R. P., White, W., Teagarden, J., & Lane, J. (2019). A meta-analysis of self-regulated strategy development reading interventions to improve the reading comprehension of students with disabilities. *Reading & Writing Quarterly, 35*(4), 339–353. https://doi.org/10.1080/10573569.2018.1545616

Schlosser, R. W., & Sigafoos, J. (2008). Identifying 'evidence-based practice' versus "empirically supported treatment". *Evidence-based Communication Assessment and Intervention, 2*(2), 61–62. https://doi.org/10.1080/17489530802308924

Schlosser, R. W., & Sigafoos, J. (2009). Navigating evidence-based information sources in augmentative and alternative communication. *Augmentative and Alternative Communication, 25*(4), 225–235. https://doi.org/10.3109/07434610903360649

Seven, Y., Hull, K., Madsen, K., Ferron, J., Peters-Sanders, L., Soto, X., Kelley, E. S., & Goldstein, H. (2020). Classwide extensions of vocabulary intervention improve learning of academic vocabulary by preschoolers. *Journal of Speech, Language, and Hearing Research, 63*(1), 173–189. https://doi.org/10.1044/2019_JSLHR-19-00052

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research (NCER 2015–002)*. National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Spencer, E., Goldstein, H., Sherman, A., Noe, S., Tabbah, R., Ziolkowski, R., & Schneider, N. (2012). Effects of an automated vocabulary and comprehension intervention: An early efficacy study. *Journal of Early Intervention, 34*(4), 195–221. https://doi.org/10.1177/1053815112471990

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*(2), 213–230. https://doi.org/10.1016/j.jsp.2013.12.002

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*(3), 142–151. https://doi.org/10.1080/17489530802505362

What Works Clearinghouse. (2020). *Works Clearinghouse Standards Handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool–Second Edition*. Harcourt Assessment.

Wood, S. G., Moxley, J. H., Tighe, E. L., & Wagner, R. K. (2018). Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities, 51*(1), 73–84. https://doi.org/10.1177/0022219416688170

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales* (5th ed.). Pearson.

Zucker, T. A., Carlo, M. S., Landry, S. H., Masood-Saleem, S. S., Williams, J. M., & Bhavsar, V. (2019). Iterative design and pilot testing of the developing talkers tiered academic language curriculum for pre-kindergarten and kindergarten. *Journal of Research on Educational Effectiveness, 12*(2), 274–306. https://doi.org/10.1080/19345747.2018.1519623

Zucker, T. A., Carlo, M. S., Montroy, J. J., & Landry, S. H. (2021). Pilot test of the Hablemos Juntos tier 2 academic language curriculum for Spanish-speaking preschoolers. *Early Childhood Research Quarterly, 55,* 179–192. https://doi.org/10.1016/j.ecresq.2020.11.009

Appendix (p. 1 of 3)

Three Sets of SAS Programming Code

First set of programming code: SAS code used to calculate an average PoGO value for each SCED study.

```
data spencer2;
input partid PoGO EV;
PoGO = PoGO / 100;
datalines;
1 33.33333333 0.0086805555
2 50 0.0138888888
3 46.875 0.007672991
4 58.82352941 0.018496264
5 44.44444444 0.011766975
6 14.70588235 0.010886750
7 72.22222222 0.014660493
8 36.11111111 0.015817901
9 27.77777778 0.021604938
;
proc mixed data = spencer2 covtest;
class partid;
model PoGO = / solution ddfm = bw notest clm;
```

Appendix (p. 2 of 3)

Three Sets of SAS Programming Code

```
random int / sub = partid;
repeated / group = partid;
parms (0.1)
(0.0086805555) (0.0138888888) (0.007672991) (0.018496264) (0.011766975)
(0.010886750) (0.014660493) (0.015817901) (0.021604938)
/ eqcons = 2 to 10;
run;
```

Second set of programming code: SAS code used to calculate an average PoGO value for all SCDs, all GEDs, and all studies.

```
data all;
input studyid partid PoGO EV design augment;
PoGO = PoGO / 100;
datalines;
1 1 42.33 0.0031427236 0 0
2 1 47.02 0.01030225 0 0
3 1 31.5357 0.00343255 0 0
4 1 47.37 0.0054493924 0 0
5 1 68.17 0.0022505526 0 1
5 2 36.31 0.0032012964 0 0
6 1 42.23 0.01014049 0 1
6 2 28.54 0.0052519009 0 0
7 1 49.186 0.01022925 1 0
8 1 28.571 0.002979751 1 0
9 1 33.53 0.00194022 1 0
10 1 47.166 0.002638063 1 0
11 1 14.35 0.001030567 1 0
12 1 33.116 0.00097471 1 0
13 1 15.161 0.00128647 1 0
14 1 27.5 0.004287358 1 0
15 1 33.601 0.000660721 1 0
16 1 41.743 0.023705124 1 0
;
data groupstudies;
set all;
if design = 1;
proc mixed data = groupstudies covtest;
class studyid;
model PoGO = / solution ddfm = bw notest cl;
random int / sub = studyid;
repeated / group = studyid;
parms (0.1)
(0.01022925) (0.002979751) (0.00194022) (0.002638063) (0.001030567)
(0.00097471) (0.00128647) (0.004287358) (0.000660721) (0.023705124)
/ eqcons = 2 to 11;
run;
data scdstudies;
set all;
if design = 0;
proc mixed data = scdstudies covtest;
class studyid partid;
model PoGO = / solution ddfm = bw notest cl;
random int / sub = studyid;
random int / sub = partid(studyid);
repeated / group = partid(studyid);
parms (0.1) (0.1)
(0.0031427236) (0.01030225) (0.00343255) (0.0054493924) (0.0022505526)
(0.0032012964) (0.01014049) (0.0052519009)
/ eqcons = 3 to 10;
run;
data studies;
set all;
```

Appendix (p. 3 of 3)

Three Sets of SAS Programming Code

```
proc mixed data = all covtest;
class studyid partid;
model PoGO = / solution ddfm = bw notest cl;
random int / sub = studyid;
random int / sub = partid(studyid);
repeated / group = partid(studyid);
parms (0.1) (0.1)
(0.01022925) (0.002979751) (0.00194022) (0.002638063) (0.001030567)
(0.00097471) (0.00128647) (0.004287358) (0.000660721) (0.023705124)
(0.0031427236) (0.01030225) (0.00343255) (0.0054493924) (0.0022505526)
(0.0032012964) (0.01014049) (0.0052519009)
/ eqcons = 3 to 20;
run;
```

Third set of programming code: SAS code used to conduct a moderator analysis.

```
data all;
input studyid partid PoGO EV design augment;
PoGO = PoGO / 100;
datalines;
1 1 42.33 0.0031427236 0 0
2 1 47.02 0.01030225 0 0
3 1 31.5357 0.00343255 0 0
4 1 47.37 0.0054493924 0 0
5 1 68.17 0.0022505526 0 1
5 2 36.31 0.0032012964 0 0
6 1 42.23 0.01014049 0 1
6 2 28.54 0.0052519009 0 0
7 1 49.186 0.01022925 1 0
8 1 28.571 0.002979751 1 0
9 1 33.53 0.00194022 1 0
10 1 47.166 0.002638063 1 0
11 1 14.35 0.001030567 1 0
12 1 33.116 0.00097471 1 0
13 1 15.161 0.00128647 1 0
14 1 27.5 0.004287358 1 0
15 1 33.601 0.000660721 1 0
16 1 41.743 0.023705124 1 0
;
proc mixed data = all covtest;
class studyid partid;
model PoGO = design / solution ddfm = bw notest cl;
random int / sub = studyid;
random int / sub = partid(studyid);
repeated / group = partid(studyid);
parms (0.1) (0.1)
(0.01022925) (0.002979751) (0.00194022) (0.002638063) (0.001030567)
(0.00097471) (0.00128647) (0.004287358) (0.000660721) (0.023705124)
(0.0031427236) (0.01030225) (0.00343255) (0.0054493924) (0.0022505526) (0.0032012964) (0.01014049)
(0.0052519009)
/ eqcons = 3 to 20;
run;
```