Running Head: EFFECT SIZE

To Appear, *Journal of Speech, Language, and Hearing Research*

Effect Size for Single-Subject Design in Phonological Treatment

Judith A. Gierut, Michele L. Morrisette and Stephanie L. Dickinson

Indiana University, Bloomington

Author Note

Judith A. Gierut, Department of Speech and Hearing Sciences, Indiana University; Michele L. Morrisette, Department of Speech and Hearing Sciences, Indiana University; Stephanie Dickinson, Department of Statistics, Indiana University.

This research was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under Award Numbers DC001694, DC00433 and DC00076 (PI: Gierut), and DC00012 (PI: Pisoni). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Dan Dinnsen for comments on the manuscript, and Grace Reynolds and Maria Miller for assisting with data analyses.

Address correspondence to Judith A. Gierut, Department of Speech and Hearing Sciences, 200 South Jordan Avenue, Indiana University, Bloomington, IN 47405-7002, Email: gierut@indiana.edu

Abstract

Purpose: To document, validate, and corroborate effect size (ES) for single-subject design in treatment of children with functional phonological disorders; to evaluate potential child-specific contributing variables relative to ES; and to establish benchmarks for interpretation of ES for the population.

Method: Data were extracted from the Developmental Phonologies Archive for 135 preschool children with phonological disorders who previously participated in single-subject experimental treatment studies. Standard Mean Difference$_{\text{All with Correction for Continuity}}$ was computed to gauge the magnitude of generalization gain that accrued longitudinally from treatment for each child, with the data aggregated for purposes of statistical analyses.

Results: ES ranged from 0.09 to 27.83 for the study population. ES was positively correlated with conventional measures of phonological learning and visual inspection of learning data based on procedures standard to single-subject design. ES was linked to children's performance on diagnostic assessments of phonology, but not other demographic characteristics or related linguistic skills and nonlinguistic skills. Benchmarks for interpretation of ES were estimated as 1.4, 3.6, and 10.1 for small, medium, and large learning effects, respectively.

Conclusion: Findings have utility for single-subject research and translation of research to evidence-based practice for children with phonological disorders.

The empirical evaluation of treatment efficacy for clinical populations has often relied on single-subject experimental design (McReynolds & Kearns, 1983; see Baker & McLeod, 2011 specific to phonological treatment). Single-subject design is well suited to clinical research because the focus is squarely on the individual as opposed to the group. Single-subject design further mirrors the clinical process by documenting an individual's pretreatment or baseline performance relative to learning that takes place as a result of intervention. Moreover, single-subject design affords the flexibility to modify treatment to accommodate individual differences in learning as in the clinical setting. Despite these and other advantages (Byiers, Reichle, & Symons, 2012), single-subject design has yet to provide opportunities for meta-analyses of treatment efficacy. Meta-analyses are considered the 'gold standard' because they provide for statistical cross-comparisons of multiple studies to establish treatment efficacy and thereby, inform evidence-based practice (Dollaghan, 2007; see Law, Garrett, & Nye, 2004 specific to phonological treatment). Single-subject design, on the other hand, relies on visual inspection of individual learning curves in evaluation of treatment efficacy. Visual inspection is the evaluation of level and/or trend in a graphic display of learning (Gast & Spriggs, 2010). Visual inspection traces relative or absolute gains in learning within and/or across treatment conditions, depending on the research objective. When differential learning is observed through visual inspection and replicated, this then serves to inform treatment efficacy.

Recent innovations in single-subject design are beginning to open the door for meta-analyses of treatment efficacy. In particular, effect size (ES) has been suggested as an adjunct to visual inspection of learning data (Busk & Serlin, 1992; Durlak, 2009; Faith, Allison, & Gorman, 1996; Komrey & Foster-Johnson, 1996; Olive & Smith, 2005). ES refers to a family of indices specific to single-subject design that establish the magnitude of gain from treatment. ES is

defined as "a quantity that characterizes the degree of departure from the null state, which, in this case, is the degree to which a treatment outcome differs from zero" (Beeson & Robey, 2006: 3). ES is scale- or unit-free, which allows for cross-comparisons of individuals, populations, experimental conditions, and studies, all from the same analytic vantage. As such, ES provides an opportunity for meta-analyses of single-subject data.

ES has long been a staple of between-group designs, with Cohen's *d* (1988) being widely applied in comparisons of two independent samples. Cohen's *d* (1988) is defined as the difference between two means divided by the standard deviation. By comparison, ES for single-subject design is just beginning to be extended to clinical populations and treatment. Research in the area of aphasia (Beeson & Robey, 2006; Robey, 1994, 1998; Robey, Schultz, Crawford, & Sinner, 1999) has led the way, with complementary data emerging, for example, in treatment of autism/developmental disability (Olive & Smith, 2005), childhood apraxia of speech (Edeal & Gildersleeve-Neumann, 2011), learning disability (Swanson & Sachse-Lee, 2000), and phonological disorders in children (Gierut & Morrisette, 2011).

The initial clinical work on ES in single-subject design has identified three general areas of research need. First, computation of ES must be specific to single-subject design. This is due to differences in the underlying assumptions of within-subject versus between-group comparisons (Busk & Serlin, 1992). Most notably, for within-subject comparisons, there are time-series dependencies associated with the longitudinal collection of data from an individual, whereas for between-group comparisons, the samples are independent. Second, benchmarks for interpretation of ES must be developed, again specific to single-subject design. Benchmarks provide a point of reference from which to descriptively characterize the magnitude of gain associated with a given ES value. For comparisons of two independent samples, Cohen (1988)

defined such benchmarks, where *d* values of 0.2, 0.5, and 0.8 are conventionally interpreted as small, medium, and large effects, respectively. Cohen (1988) coined the terms 'small', 'medium' and 'large' effects as rule of thumb descriptors, handy for interpretation but arbitrary in expression. Cohen (1988) further cautioned that benchmarks are only applicable to the specific lines of inquiry for which they were originally intended. Consequently, Cohen's benchmarks may not be applicable to single-subject design, longitudinal, or clinical research (Durlak, 2009; Faith et al., 1996). Third and relatedly, benchmarks for interpretation of ES must be specific to the population of study (Beeson & Robey, 2006). Population-specific benchmarks allow for apples-to-apples comparisons because the magnitude of gain achieved in treatment of one disorder may not be the same as another.

This paper begins to address these general issues in the context of treatment of phonological disorders in children. Phonological disorders represent a subgroup of speech sound disorders that are functional in nature and affect the linguistic structure, organization, representation, and/or rule-governed use of the sounds of language (Gierut, 1998). By way of background, we describe one ES computation that is well suited to single-subject studies of phonological treatment. A review of the literature that applied this computation in evaluation of phonological treatment follows. The available work sets the stage for the present study of ES in 135 children who previously participated in single-subject research on phonological treatment.

Standard Mean Difference

There are a variety of ES computations that may be applied to single-subject data, each with different merits, utility, and power (Busk & Serlin, 1992; Campbell & Herzinger, 2010; Faith et al., 1996; Hoyle, 1999; Olive & Smith, 2005). The selection of which ES computation

to use is dictated by the research question, experimental design, and dependent variable of interest (Durlak, 2009).

With respect to the research question, the study of phonological treatment has taken a three-pronged approach that aligns with the definitions of treatment effectiveness, efficiency, and effects (Olswang, 1998). Studies that introduce novel methods of treatment document effectiveness (e.g., Dean, Howell, Waters, & Reid, 1995; Miccio & Elbert, 1996); those that compare different protocols of treatment establish efficiency (e.g., Hesketh, Adams, Nightingale, & Hall, 2000; Powell, Elbert, Miccio, Strike-Roussos, & Brasseur, 1998); and, others that report differential learning as a consequence of instruction document treatment effects (e.g., Gierut & Morrisette, 2012a, 2012b; Tyler & Figurski, 1994). Of these strands of study, the literature is replete with the latter, with evidence gleaned largely from experiments that utilize the multiple baseline design.

With respect to the experimental design, the assumptions and setup of the multiple baseline have been described in detail elsewhere (Byiers et al., 2012; Gast, 2010; Hersen & Barlow, 1976; Kratochwill, 1978; McReynolds & Kearns, 1983); but, briefly, a no-treatment phase is followed by a treatment phase. The *no-treatment phase* documents baseline performance prior to treatment, with the number of baseline samples increasing as successive children enroll in a given experimental condition. The baseline is a measurement of the specific behaviors that treatment is intended to change. Baseline performance is expected to remain stable within and across children until the instatement of treatment, with stability operationalized as +/– 10% variation (McReynolds & Kearns, 1983). Baseline stability is fundamental to single-subject design because it ensures internal validity (Kratochwill, 1978). Consider, for example, that rising baselines might reflect extraneous influences of maturation; saw-tooth baselines might

suggest extraneous influences of the environment; falling baselines might imply confounds associated with invalid measurement tools (Kratochwill, 1978). Thus, baseline stability contributes to the demonstration of experimental control. The subsequent *treatment phase* is intended to shift performance from baseline in lockstep with intervention. A period no-change in baseline followed by a period of change exclusive to, and concurrent with intervention establishes a functional relationship between learning and treatment. This allows "cause and effect statements about the dependent and independent variables, statements regarding the functional relationship between treatment variables and behavioral change" (McReynolds & Kearns, 1983: 7). The logic is that the time-locked change in performance is attributable to treatment, not other extraneous variables.

As applied to phonological disorders, sounds that a child excludes from his or her phonemic inventory are often the targets of treatment in the multiple baseline design. This is for two reasons. Children with phonological disorders have reduced phonemic inventories due to constraints on the phonotactics of their grammar and often produce only nasals, stops, and glides to contrast meaning differences among words. Treatment of sounds excluded from the phonemic inventory thereby addresses a core source of children's errors, with an emphasis on singletons. Moreover, by targeting sounds excluded from the phonemic inventory, baseline performance is guaranteed to be stable, near 0% accuracy within the allowable range of variation. Treatment is then applied to shift production accuracy from a zero-change state. A functional relationship is established when gains in production of sounds excluded from the phonemic inventory coincide with the delivery of treatment. Each child enrolled successively in the multiple baseline design allows for replication of the functional relationship between learning and treatment.

Finally, with respect to the dependent variable, accuracy of production of the treated sound in treated words is documented session-by-session in phonological treatment. These data are used to advance a child through the instructional protocol and to demonstrate that learning indeed occurred. However, an ultimate gauge of phonological treatment efficacy is system-wide generalization (Gierut 1998; Powell, 1991; Tyler & Figurski, 1994). System-wide generalization is defined as the extension of accurate production to treated and untreated erred sounds (excluded from the phonemic inventory) across phonetic contexts and lexical items. The overarching goal is to optimize learning by promoting broad system-wide gains in a child's phonology. Generalization thus serves as the primary measure of learning. Generalization is sampled longitudinally throughout the course of treatment, but is a process independent of session-by-session performance in treatment (McReynolds & Kearns, 1983).

One ES computation that fits the intent, design, and dependent variable of phonological treatment is Standard Mean Difference$_{\text{All with Correction for Continuity}}$ (Busk & Serlin, 1992; see Gierut & Morrisette, 2011 for phonological disorders). This is essentially a single-subject analog to Cohen's *d*. Standard Mean Difference does not assume normality, equal variance, or serial independence of data. As such, it is a good match to single-subject design. It also is a computation of choice in evaluations of treatment for clinical populations generally (e.g., Beeson & Robey, 2006; Olive & Smith, 2005; Thompson, den Ouden, Bonakdarpour, Garibaldi, & Parrish, 2010). Standard Mean Difference is a nonregression technique that is thought to be more conservative than regression measures, which tend to overestimate learning effects (Busk & Serlin, 1992). Its further appeal lies in ease of computation and interpretability given its similarity to Cohen's *d* (Olive & Smith, 2005). The latter are relevant considerations if ES is to be employed by practicing clinicians in applied settings.

To compute Standard Mean Difference, the mean accuracy of production is established for repeated baselines. The mean accuracy of *all* generalization data collected longitudinally over the course of treatment is likewise computed. The difference between means is then calculated and forms the numerator of the operation; hence, the term Standard Mean Difference$_{All}$ (Olive & Smith, 2005). The standard deviation of the baseline pooled for the study population forms the denominator. Pooling baseline data for the study population to obtain the standard deviation was recommended by Glass (1977; Busk & Serlin, 1992) as a way to handle potential 0% baseline variability within-subject. Consider that if the baseline is perfectly stable with no variability, it is not possible to compute a standard deviation for use as the denominator. Moreover, if baseline performance is near floor (as expected for sounds excluded from children's phonemic inventories), and if there is small (non-zero) variation (as is called for in the multiple baseline design), measures of standard deviation would be unstable. By pooling baseline data, variability in performance is established specific to the study population and further serves as a correction for continuity (cf. Beeson & Robey, 1996; Busk & Serlin, 1992 for alternatives). When the difference between the mean baseline and generalization is divided by the pooled standard deviation, the resulting ES value (*d*) is the Standard Mean Difference. Standard Mean Difference is essentially a comparison of the means of two distributions: baseline relative to generalization.

ES in Phonological Treatment

To our knowledge, four multiple baseline studies of phonological treatment have reported ES based on computation of Standard Mean Difference (Gierut & Morrisette, 2011, 2012a, 2012b, 2014). Table 1 shows the sample size, stimulus conditions of treatment, and ES values based on the magnitude of generalization from treatment associated with each study. Studies

were similar in many respects. Each enrolled a homogeneous group of preschool children who met the same inclusionary and exclusionary criteria for participation. All children presented with severely reduced phonemic inventories, and all received treatment on accurate production of a sound excluded from the inventory. Across studies, the stimulus conditions to promote phonological generalization were thematically related (e.g., teach frequent vs. infrequent words or early vs. late acquired words). Likewise, generalization was the primary measure of learning, with the same structured probe being used to sample treated and untreated sounds excluded from the phonemic inventory. Further, each study used ES to corroborate visual inspection of longitudinal generalization data, where visual inspection was defined as absolute differences in level of performance relative to baseline within and across experimental conditions. Several observations can be gleaned from this set of preliminary studies to inform ES for phonological treatment. In turn, the observations revealed gaps that warrant attention as motivation for the present study.

Insert Table 1 about here

One observation is that ES varied across children with phonological disorders. This can be seen in Table 1, for example, where ES reportedly ranged from 2.6-16.58 for the population of study. Such variation is not unexpected, given well-documented cases of individual differences in phonological acquisition (Vihman, Ferguson, & Elbert, 1986) and treatment (Dean et al., 1995). This notwithstanding, the scope of variability in ES and the upper and lower bounds are not known. This information is needed to better understand the extent of generalization that may reasonably be expected as a consequence of phonological treatment.

Another observation is that ES varied within and across studies. Table 1 shows, for example, that when age-of-word-acquisition was manipulated (Gierut & Morrisette, 2012a),

sounds taught in late acquired words resulted in greater ES values than in early acquired words, i.e., 11.41-16.58 versus 2.81-3.66, respectively. Further, late acquired words that were infrequently occurring in the language yielded greater ES values than late acquired words that were frequently occurring, i.e., 16.58 versus 11.41, respectively. ES values may thus be used to establish relative rankings of the efficacy of different experimental conditions. However, what is not known is whether simple differences in raw ES values reflect substantive differences in children's generalization learning. This bears on Bothe and Richardson's definition of practical versus clinical significance (2011; cf. Bain & Dollaghan, 1991; Durlak, 2009; Jacobson & Truaz, 1991; Rosenthal & Rosnow, 2008 for use of complementary terms). Practical significance relies on ES to establish the absolute magnitude of gain from treatment, whereas clinical significance relies on clinical measures interpretable to practitioners to establish the same effects. Thus far, the relationship between ES and conventional measures of phonological learning remains unknown, but is needed in the translation of research to evidence-based practice.

A related observation is that, while relative rankings of ES might inform the results of treatment, they lack descriptive interpretation. It is not known, for example, whether the ES values shown in Table 1 correspond to small, medium, or large learning effects, using the rule of thumb descriptors coined by Cohen (1988). It is possible that different ES values actually fall into the same interpretive category. Therefore, it is necessary to document ES for a large cohort of children with phonological disorders and to establish benchmarks statistically based on the distribution that obtains.

Two other gaps are worth mentioning. ES has been used as an analytic supplement to visual inspection of generalization. Presumably, greater generalization will correspond to greater ES values; however, the association between ES and visual inspection of learning data has not

been validated for phonological treatment. Further, ES has been examined only in connection with phonological treatment, consistent with the logic of a functional relationship in single-subject design. Faith and colleagues (1996) suggest that this be verified by assessing possible contributing variables relative to ES. For the study of phonological disorders, it has yet to be determined whether factors other than treatment uniquely impact children's generalization. This is relevant because a host of child-specific factors have been implicated as causal, co-occurring, or contributing to the disorder (e.g., recurrent otitis media, Miccio, Gallagher, Grossman, Yont, & Vernon-Feagans, 2001; word learning, Shriberg & Kwiatkowski, 1994; phonological working memory, Shriberg, Lohmeier, Campbell, Dollaghan, Green, & Moore, 2009). It is necessary to likewise establish whether such factors affect generalization indexed by ES.

This paper begins to address these questions in a retrospective examination of ES derived from generalization learning by 135 children with phonological disorders who were previously enrolled in single-subject multiple baseline studies of treatment. The purpose is five-fold: (1) to document ES for phonological treatment, (2) to determine the relationship between ES and conventional measures of phonological learning, (3) to verify ES as a complement to visual inspection of generalization data, (4) to examine possible factors contributing to ES, and (5) to suggest benchmarks for interpretation of ES specific to the population of children with phonological disorders.

Methods

Data for analysis were drawn from the Developmental Phonologies Archive housed at Indiana University (Gierut, 2008b). This is an electronic compendium of descriptive and experimental results of clinical treatment studies enrolling children with phonological disorders. Inclusionary and exclusionary criteria for enrollment, participant characteristics, descriptive and

experimental methods, and the archive have all been described in detail elsewhere (Gierut, 2008a, b; Gierut, Morrisette, & Ziemer, 2010); only information central to the present study is outlined herein. We begin with the rationale and data for inclusion. This is followed by description of the study population, treatment protocol, stimulus conditions, generalization measure, fidelity, and reliability. Procedures for data analyses and autocorrelation of the data are also reported.

Rationale and Data for Inclusion

Data from the Developmental Phonologies Archive were amenable to an evaluation of ES for several reasons. Inclusionary and exclusionary criteria for participation were the same for all children, with a core battery of diagnostic tests used to establish eligibility. Phonological treatment was administered in the same way to all children in accord with a protocol that capped time in treatment. Treatment was uniformly directed at improving accuracy of production of sounds excluded from a child's pretreatment phonemic inventory, with the focus on singletons. Generalization was the primary measure of learning for all children. Generalization was based on a child's performance on a structured probe that was also identical for all children. The probe sampled treated and untreated sounds excluded from each child's phonemic inventory across contexts and in multiple lexical items from which it was possible to compute percent accuracy of production. Generalization likewise focused on gains in production of singletons. Frequency of probe administration followed a comparable schedule averaging every third session. The duration of probe administration remained the same for all children. Generalization was tracked during treatment, with follow-up after treatment was withdrawn for insight to maintenance of learning effects. Uniformity of the archival data thus offered an opportunity to establish ES for a relatively homogeneous group of children exposed to similar conditions in inducing and monitoring phonological generalization. Comparability of the data lent a further advantage in

that ES could be computed specifically for each child. This circumvented known limitations associated with estimating ES through extrapolation of data (Glass, 1977; Olive & Smith, 2005).

At the outset of the present study, data from 251 children were available in the archive. For a child's data to be included in the computation of ES, five conditions had to be met.

(1) A child had to receive phonological treatment, excluding 39 cases enrolled for descriptive purposes only. Children excluded for this reason contributed just one phonological sample and were not followed longitudinally. Therefore, it was not possible to assess generalization over time.

(2) A child had to complete the treatment protocol, excluding 11 cases of attrition. Attrition is a known threat to the internal and external validity of single-subject design (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005): Data sets are incomplete and too few; truncated data preclude the evaluation of learning in a manner consistent with other participants; sporadic attendance compromises learning and contaminates the interpretation of treatment effects.

(3) A single-subject multiple baseline design had to be employed in treatment given its fit to Standard Mean Difference, excluding 57 cases. Children excluded for this reason had been enrolled in multiple probe or alternating treatments designs.

(4) Baseline production of sounds excluded from the phonemic inventory had to remain stable, within +/- 10% variation from the mean to avoid spurious cases of spontaneous improvement, excluding 0 cases. Baseline stability ensures internal validity and experimental control in single-subject design (Kratochwill, 1978).

(5) A child had to evidence gains in production of sounds excluded from the pretreatment phonemic inventory relative to baseline, excluding 9 cases. Children excluded for this

reason showed generalization, but only in production of consonant clusters. The focus herein was on generalization to singletons, with clusters beyond the scope of study. When these five conditions were applied, data from 135 children remained for analysis.

Data from 132 of 135 children (98%) in the study population were reported in the published literature, with primary sources available at www.indiana.edu/~sndlrng. Data for the remaining 3 children were presented at a professional meeting (Morrisette, Hoover, & Gierut, 2012). Only participants of a given study who met the aforementioned conditions for data inclusion entered into the ES analyses. It is of further mention that 132 of 135 children (98%) maintained gains in production of sounds excluded from the phonemic inventory over multiple sampling points in time, suggesting that the generalization effects were not transient.

Characteristics of the Study Population

As noted, children met the same inclusionary and exclusionary criteria for enrollment. Specifically, children were monolingual English speakers between the ages of 3 and 7. They performed at or below 1 standard deviation from the mean on the *Goldman-Fristoe Test of Articulation* (*GFTA*, Goldman & Fristoe, 1986, 2000) and produced at least 6 sounds in error across contexts on this measure. In addition, they performed within typical limits on a battery of diagnostic tests that included hearing acuity, oral motor structure-function, vocabulary, language, cognition, and working memory (Gierut, 2008b). They were preliterate based on parent report and not enrolled concurrently in any other type of intervention program. Additional case information was obtained, but not used to determine eligibility for participation.

In all, there were 89 boys and 46 girls in the study population; their average age was 4 years, 5 months (*SD* = 10; range 36-93). By parent report, 89% of the children produced their first words on track developmentally; 50% had a family history of speech-language-hearing

disorders; and 42% had a prior history of otitis media. As will be seen, demographic and diagnostic entry test results will be relevant herein to the evaluation of contributing factors relative to ES.

In addition to standardized testing, detailed clinical and linguistic analyses of the phonologies of each child were developed and available in the Developmental Phonologies Archive. On the clinical side, Percent Consonants Correct-Revised (PCC-R; Shriberg, Austin, Lewis, McSweeny, & Wilson, 1997) was computed as an index of severity and Proportion of Whole Word Proximity (PWP; Ingram & Ingram, 2001), as a measure of the preservation of word shape and consonantal accuracy. Both analyses were based on 50-word samples, with scores established at pretreatment and again immediately following completion of treatment. Reliability in calculation of PCC-R and PWP was established by two independent judges for 10% of the archival data, with 99% agreement. Cohen's kappa coefficient (κ) was also applied to control for chance agreements in scoring. The obtained kappa value was .99, which Landis and Koch (1977) describe as almost perfect agreement.

For the study population, the proportional mean PCC-R score was .49 (*SD* = .14, range = .10-.78), which corresponded to the severity descriptor 'severe', whereas proportional mean PWP was .73 (*SD* = .09, range = .45-.91). As will be seen, children's PCC-R and PWP scores, as conventional clinical measures, were relevant to establishing the validity of ES as an index of generalization gain.

On the linguistic side, conventional independent and relational descriptions (Dinnsen, 1984; Stoel-Gammon, 1985) were developed, with particular attention to sounds excluded from children's pretreatment phonemic inventory. Sounds excluded from the phonemic inventory were relevant because these were monitored for all children as a reflection of learning. Sounds

excluded were determined from the results of an established structured probe (Gierut, 2008b). The probe was administered as a picture-naming task and consisted of 293 words that sampled each target English consonant in each relevant word position in multiple exemplars. Any given consonant was sampled in at least 17 mono- and bimorphemic words, with opportunities for minimal pairs. The probe was used exclusively as a test measure. Sounds excluded from the phonemic inventory were identified using probe data obtained pretreatment following established criteria (Gierut, Simmerman, & Neumann, 1994). Namely, these were sounds produced with near 0% accuracy across phonetic contexts and also, these same sounds were never used by a child to mark meaning distinctions in minimal pairs.

For the study population, children's phonemic inventories consisted of an average of 14 phonemes (*SD* = 2.93, range = 7-20), with 9 sounds excluded from the repertoire (*SD* = 2.93, range = 3-16). Sounds excluded from the phonemic inventory were primarily velar stops, fricatives, affricates, and liquids. Of relevance herein, children's pretreatment phonemic inventories delineated the sounds to be evaluated for magnitude of generalization gain in computation of ES.

Treatment Protocol

The treatment protocol was standard to all children, as in the appendix. A certified speech-language pathologist provided individualized instruction to each child in 1-hr sessions, 3 times weekly. Treatment consisted of two phases: imitation followed by spontaneous production of the treated sound in treated stimuli. Treated sounds were specific to a given child's presenting errors and treated stimuli, specific to the experimental questions of interest. Treated stimuli were used only for instruction and were never tested as evidence of generalization. During the imitation phase, a child was provided with a model of the treated sound in treated stimuli and

instructed to repeat, with corrective feedback provided. Imitation continued for a total of 7 sessions or until a child achieved 75% accuracy of production of the treated sound in treated stimuli over two consecutive sessions, whichever occurred first. Treatment then shifted to the spontaneous phase, where a child produced the treated sound in treated stimuli without benefit of a preceding model; as in imitation, corrective feedback was again provided. The spontaneous phase continued for a total of 12 sessions or until a child achieved 90% accuracy of production of the treated sound in treated stimuli over 3 consecutive sessions, whichever occurred first. The appendix details the treatment sequence, citing the baseline, instructions, criteria for advancement, and schedule of probe administration as the measure of generalization. Fidelity in administration of the treatment protocol was documented for 10% of 135 children. An independent observer used an established checklist procedure (Gierut, 2008a) to ensure that the protocol was administered as prescribed in the appendix. Fidelity was judged to be 100%.

For the study population, children received an average of 14 sessions of treatment (*SD* = 4.87, range = 5-19). They required 5 mean sessions (*SD* = 1.94, range = 2-7) to complete the imitation phase, and 8 mean sessions (*SD* = 3.52, range = 3-12) to complete the spontaneous phase of the protocol. During imitation, children achieved, on average, 82% maximum accuracy of production (*SD* = 13.51, range = 24-100) of the treated sound in treated stimuli. During the spontaneous phase, they achieved, on average, 94% maximum accuracy of production (*SD* = 7.48, range = 56-100) of the same stimuli. These data demonstrated that treatment led to improved accuracy of production as the springboard for subsequent generalization. As will be seen, children's time and performance in treatment will be evaluated to determine the extent to which these factors potentially contributed to ES.

Stimulus Conditions of Treatment

The stimulus conditions of treatment were the one element that varied across the study population, but were thematically related in that treated sounds (e.g., markedness relationships, Gierut, 2007) or treated stimulus words (e.g., word frequency, Morrisette & Gierut, 2002) were manipulated. The effects of such manipulations were evaluated previously using visual inspection of generalization data. As noted above, visual inspection evaluated absolute levels of generalization that obtained within and across experimental conditions. Replications of differential generalization were then used to discern the relative treatment effects of given conditions. General findings have been collectively summarized in the literature (Gierut, 2001, Table 1; 2007, Table 2; see www.indiana.edu/~sndlrng for primary sources). For the study population, 63 of 135 children were previously assigned to an experimental condition that resulted in relatively greater generalization based on visual inspection of learning data. The remaining 72 children were previously assigned to a condition that resulted in relatively less generalization. As will be seen, differential generalization in the study population will be relevant to establishing ES as a statistical complement to visual inspection of single-subject data.

Generalization as the Measure of Learning

Generalization to sounds excluded from a child's phonemic inventory relative to baseline was the primary source of data that entered into computation of ES in this study. Recall that generalization is the transfer of learning from treatment to untreated stimuli and is independent of performance session-by-session during instruction (McReynolds & Kearns, 1983). Generalization was measured based on a child's performance on the aforementioned structured probe. The probe was reserved exclusively as a test measure and was never employed during treatment. The probe was administered longitudinally, with samples obtained before, during, and immediately upon completion of treatment. These data entered into the computation of ES to

determine the magnitude of generalization that occurred as a function of treatment. Additional samples were collected after withdrawal of treatment, continuing to approximately 8 weeks. These data were for descriptive purposes only in evaluation of maintenance.

Number of probe samples. Probes obtained before treatment established baseline performance. In keeping with the multiple baseline design, the number of baselines was incremented by 1 as successive children enrolled in a given experimental condition. As such, the number of baselines varied across children. For the study population, the average number of baselines was 3 (*SD* = .83; range = 1-5).

Probes administered during, and upon completion of treatment informed the functional relationship between treatment and generalization relative to baseline. These probes were administered following a variable ratio schedule averaging three sessions (Appendix); accordingly, the number of probes in treatment likewise varied across children. For the study population, the average number of probes collected during treatment was 7 (*SD* = 2.96; range 2-12).

Probes administered after withdrawal of treatment informed maintenance of generalization. For the study population, the average number of probes in withdrawal was 2 (*SD* = .39; range = 1-4). Thus, an average of 12 total probes were administered to each child (*SD* = 3.14; range = 5-18).

Reliability of probe transcription. Throughout, a child's probe responses had been digitally recorded. Subsequently, a trained listener phonetically transcribed the data. Reliability of phonetic transcription was established by a second independent listener, who was naïve to the children and experimental questions of interest. Point-to-point agreement in consonant transcription was established for 10% of longitudinal probe data. Standard procedures for

establishing interjudge transcription reliability were used (McReynolds & Kearns, 1983; Shriberg & Lof, 1991). Mean agreement was 92% based on 40,240 segments transcribed (*SD* = 3; range = 82-98).

Analysis of probe data. The transcribed longitudinal probe data had been entered into the Developmental Phonologies Archive from which accuracy of production of sounds excluded from a child's phonemic inventory was computed. Production accuracy was determined for each child and each probe sample. Recall that the probe evaluated all English consonants, however, only sounds excluded from a child's phonemic inventory with stable baselines were monitored for generalization; hence, only relevant probe words were examined for accuracy. For the study population, a total of 263,509 probe words were evaluated for accuracy. Each child contributed a mean of 1,952 probe words (*SD* = 338; range = 1,096-2,809), with approximately 163 words evaluated for accuracy at each probe point. These were the longitudinal data used to compute ES.

Computation of ES. Each child's probe data was evaluated independently to arrive at an ES based on Standard Mean Difference. Recall that the formula computes the difference between a child's mean baseline performance averaged over successive samples and mean generalization during treatment, likewise averaged over successive samples. The difference between means is then divided by the standard deviation of the baseline for the population to derive an ES value. For the study population of 135 children, the standard deviation of the baseline was .02 (range = .00-.08), where the baseline reflected accuracy of production of singleton sounds excluded from the phonemic inventory prior to treatment. This formula was applied to the longitudinal probe data from each child to yield an ES value. Data from the study population were then aggregated in statistical analyses.

Autocorrelation of data. Due to the time-series nature of single-subject design, first lag autocorrelations were computed to assess bias in the aforementioned data. Autocorrelations determine the extent to which successive data points are correlated. If autocorrelation coefficients are positive, this suggests liberally biased errors in the sample; if negative, this suggests conservatively biased errors (Crosbie, 1987). In this study, autocorrelations were calculated independently for probe data obtained at baseline and through to completion of treatment (i.e., generalization) as each was integral to ES. For baseline data, autocorrelation coefficients were available for 76 of 135 children. (It should be noted that the statistical run outright eliminates constant values and/or fewer than two data points; hence, baseline data were trimmed accordingly.) The resulting mean autocorrelation coefficient was −.30 (*SD* = .25; range = −.73-.25), suggesting that baseline data were conservatively biased. For probe data collected in treatment, autocorrelation coefficients were available for 116 of 135 children. The mean autocorrelation coefficient was .01 (*SD* = .29; range = −.69-.66), suggesting that positive autoregressive effects on generalization data were minimal.

Results and Discussion

Results are organized to address five overarching questions associated with the application of ES for single-subject design in treatment of children with phonological disorders. The distribution of ES for the study population of 135 children is reported using descriptive statistics. ES data were then submitted to a series of inferential analyses to establish validity relative to other measures of learning, corroboration relative to visual inspection of learning data, and the potential influence of other contributing variables. These analyses laid the groundwork for estimating benchmark categories to differentiate small, medium, and large effects for use as rule of thumb descriptors in interpretation of ES for phonological disorders. Thus, ES is

considered from the vantage of description, validation, corroboration, contributing influence, and interpretation.

Distribution of ES

Figure 1 presents a histogram of the distribution of raw ES for the study population. Raw ES values ranged from 0.09 to 27.83 and were right-skewed (skewness = 2.73). Consequently, the raw data were log transformed to better approximate a normal distribution for purposes of statistical analyses. A natural log scale was used, where x = ln(1 + ES). Figure 2 plots the distribution of log ES, where skewness = 0.46. All descriptive and inferential statistics were based on these transformed data. To aid interpretation, means and confidence intervals (CI) from the log scale were uniformly back-transformed ($ES = e^x - 1$) and are reported herein as raw ES with corresponding CIs. From Figure 2, the mean log ES was 1.54 (*SD* = .62), with 95% CI [1.43, 1.64], which corresponded to the mean raw ES of 3.66, 95% CI [3.20, 4.18] for children with phonological disorders.

Insert Figures 1 and 2 about here

The ES measures were evaluated for possible regression to the mean. Regression to the mean is a statistical phenomenon for difference measures whereby random variation gives the appearance of 'real' change. If regression to the mean were operative, children with greater performance at baseline would show less generalization than those with poorer baseline scores. Analyses of such data would thereby result in negative correlations between baseline and generalization. A Pearson correlation was performed to establish the degree to which ES was related to baseline performance. The correlation was close to zero and not statistically significant, $r(133) = .07$, $p = .43$, thereby abating concerns about regression to the mean.

Several points about the descriptive results are worth highlighting. One observation is that ES values for children with phonological disorders were similar to those reported in single-subject treatment of other linguistic disorders. In aphasia, for example, ES in treatment and generalization have reportedly ranged from 2.01-23.92 and 0-13.28, respectively (Robey et al., 1999; Thompson et al., 2010). By comparison, single-subject treatment of nonlinguistic skills has resulted in smaller ES values. For example, ES in treatment of developmental disability ranged from 0-3.0 (Olive & Smith, 2005) and learning disabilities, 0.58-1.13 (Swanson & Sachse-Lee, 2000). This aligns with the recommendation that ES be determined specific to behaviors of interest and populations of study (Beeson & Robey, 2006).

Another observation is that ES values for single-subject design have been generally greater than those associated with between-group designs, where Cohen (1988) cites benchmark values of 0.2-0.8 for small-to-large effects. This highlights the necessity of using ES calculations specific to the nature of the experimental design.

A further observation is that the 95% CI [3.20, 4.18] for mean raw ES in phonological disorders was narrow. This is relevant because the CI provides an estimate of generalizability to the broader population of interest: The narrower the CI, the more likely the ES values are generalizable (Law et al., 2004). The suggestion is that ES values obtained herein are representative of the gains to be expected in treatment of children with phonological disorders generally.

A final observation relates to regression to the mean and the protections offered by single-subject design. There is consensus in the literature (Barnett, van der Pols, & Dobson, 2004; Linden, 2013; Nesselroade, Stigler, & Baltes, 1980) that evaluations of change derived from a single baseline sample compared to a single posttreatment sample are vulnerable to

regression to the mean; however, regression to the mean is minimized when baseline samples are repeated and averaged. The relevance here is that the multiple baseline design necessitates successive baselines and moreover, computation of Standard Mean Difference is based on the averages of two distributions, baseline and generalization. Together, these inherent design features guard against spurious interpretations of change. Linden (2013: 6) suggests that "designing interventions to mitigate the effects of RTM [regression to the mean] is a preferred strategy to retrospectively estimating the extent to which RTM may explain any observed treatment effect." Nesselroade and colleagues (1980) further suggest that the study of child development itself wards against regression to the mean. The reason is that development follows a characteristically complex heterochronic trajectory that cannot be adequately handled by standard regression models.

ES as a Valid Index of Learning

Analyses were completed to establish ES as a valid measure of phonological learning. The intent was to determine whether ES correlated with conventional measures that have been previously used in evaluation of phonological treatment. Three conventional measures were selected as representative of the single-subject literature. These included the difference in (1) PCC-R scores pre- to posttreatment (following e.g., Hesketh et al., 2000; Tyler & Figurski, 1994); (2) PWP scores pre- to posttreatment (following e.g., Ingram, 2002; Ingram & Ingram, 2001); and (3) production accuracy pre- to posttreatment (following e.g., Miccio & Elbert, 1996; Powell et al., 1998). Notice that each conventional measure of phonological learning involves a two-shot comparison of a child's performance on a single pretreatment versus a single posttreatment sample, represented as a difference score. This contrasts with ES, which relies on averaged data from multiple samples that accrue longitudinally in baseline versus treatment.

Insert Table 2 about here

Difference scores were computed for each of the aforementioned conventional measures for each child of the study population. To illustrate, if a given child had a proportional PCC-R score of .55 pretreatment and a corresponding score of .63 posttreatment, the difference (gain) in PCC-R was .08. If that same child had a proportional PWP score of .75 pretreatment and .82 posttreatment, the difference (gain) was .07. Similarly, if the child's production accuracy at pretreatment was proportionally .04 and at posttreatment .76, the difference (gain) was 72. The resulting difference scores were normalized when skewed and submitted to Pearson correlation analyses relative to ES, as in Table 2. PCC-R and PWP difference scores were normally distributed, where skewness = .71 and –.24, respectively. Production accuracy difference scores were log transformed before analysis, where skewness = 2.57; hence, back-transformed values are reported in Table 2. The data in Table 2 show that ES was positively correlated with each measure of phonological learning, all *p*s $\leq$ .001. This demonstrates that ES converged with other conventional measures that have been used previously in the single-subject literature to evaluate phonological treatment.

The correlation of ES with PCC-R and PWP scores is of particular interest from the vantage of translational research. The reason is that ES reflects the practical significance of treatment effects, whereas PCC-R and PWP reflect the clinical significance of the same effects (Bothe & Richardson, 2011). The latter are techniques commonly used by, and readily interpretable to practicing clinicians. An implication is that experimental treatment studies that report ES may inform clinical practice because it appears that ES and clinical measures of phonological gain dovetail. The correlation of ES with pre- to posttreatment production accuracy is also worth noting given the time course involved. Because ES was based on

cumulative longitudinal data, it captured dynamic phonological learning as it unfolded over time. This contrasts with pre/post comparisons of accuracy, which yielded relatively static characterizations of learning based only on start and end points. An implication is that dynamic and static measures may offer a unified perspective on phonological learning that obtains from treatment.

Table 2 also reports the correlation between ES and maintenance of generalization. Recall that sounds excluded from children's phonemic inventories were monitored after treatment was withdrawn. Maintenance data in Table 2 were log transformed before analysis, where skewness = 1.80; hence, back-transformed values are reported. It can be seen that ES was again positively correlated with maintenance, $r(133) = .76$, $p < .001$. The greater the ES at completion of treatment, the greater the gains after treatment was withdrawn. ES appears to be predictive of continued phonological learning in the absence of intervention. Taken together, these results established the validity of ES in single-subject design given its convergence with other conventional measures that have been used previously to evaluate phonological treatment.

ES and Visual Inspection of Data

Analyses were completed to establish the degree of convergence between ES and visual inspection of generalization data. The intent was to determine whether ES aligned with differential generalization patterns previously reported in the literature for the study population. Recall that the protocol of treatment (Appendix) was constant for the 135 children of study, but the stimulus conditions manipulated in treatment differed across children. Recall too that 63 of 135 children were assigned to stimulus conditions affiliated with greater generalization as reported in the literature based on visual inspection of absolute level of performance relative to baseline. The mean back-transformed raw ES for this subgroup of children was 4.70 (95% CI

[3.84, 5.70], range = .25-27.83). The remaining 72 children were assigned to stimulus conditions associated with relatively less generalization, also based on visual inspection. Their corresponding mean back-transformed raw ES was 2.91 (95% CI [2.46, 3.43], range = .09-16.43). An independent samples t-test showed that the two subgroups were significantly different in ES, $t(133) = 3.66$, $p < .001$. Thus, ES as a statistical index converged with published reports based on visual inspection of generalization data. The finding might appear simplistic because ES was derived based on larger subsets of previously reported generalization data; however, this demonstration was necessary as a precursor to future work that might involve comparative evaluations of treatment effects within and across children and/or studies. This is a point taken up in the general discussion.

Potential Contributing Variables

Analyses were completed to establish the degree to which ES was influenced by factors other than treatment. The intent was to determine whether ES was affected by individual differences associated with children's presenting skills and/or session-by-session performance in treatment. Table 3 summarizes child demographics and performance on diagnostic tests of phonological and other related skills, which were used to establish eligibility for participation. Table 4 summarizes data from session-by-session treatment. Continuous variables were submitted to Pearson correlation analyses and binary variables to two-sample t-tests relative to ES.

Insert Tables 3 and 4 about here

Beginning with children's presenting skills, Table 3 shows that diagnostic, clinical, and linguistic assessments of phonological skills were all positively correlated with ES, all ps $\leq .02$. Children with better phonological skills on diagnostic tests at enrollment uniformly achieved

greater magnitudes of gain from treatment. By comparison, Table 3 shows that children's demographic characteristics and performance on tests of other linguistic and nonlinguistic skills were not correlated with ES, all *p*s ≥ .07. It is of note that several characteristics and skills examined herein have been previously implicated in the occurrence of phonological disorders (e.g., recurrent otitis media, Miccio et al., 2001; word learning, Shriberg & Kwiatkowski, 1994; phonological working memory, Shriberg, et al., 2009). Apparently, these same factors did not seem to differentially influence ES (see Shriberg, Kwiatkowski, & Gruber, 1994 for a similar finding). Turning to session-by-session treatment considerations, Table 4 further shows that the number of treatment sessions, accuracy of production of the treated sound in treated stimuli, and number of probe samples were not correlated with ES, all *p*s ≥ .22. This highlights the distinction between learning versus generalization (McReynolds & Kearns, 1983).

Together, the findings showed that ES was related only to children's phonological skills and not other contributing variables. ES seemed to be confined to the phonological domain and reflected change only in that domain. It is possible that this was related to homogeneity of the study population given the stringent inclusionary and exclusionary criteria for participation. It is also possible that ES lacks sensitivity in detecting the influence of child-specific or external factors (Campbell, 2004). These possibilities will be revisited in the general discussion as directions for future research.

Benchmarks for Interpretation

The final set of analyses intended to identify boundaries that define small, medium, and large learning effects for children with phonological disorders. The aim was to align ES values with corresponding descriptors to aid interpretation. The descriptors 'small', 'medium', and 'large' were borrowed from Cohen (1988) as put forth for interpretation of tests of independent

samples. Likewise, these descriptors have been accepted in interpretation of single-subject treatment studies of other clinical populations (Robey et al., 1999). For consistency with the broader literature, the same terms were adopted herein.

ES data were submitted to a k-means cluster analysis specified to isolate three groups. This technique identifies natural breaks in the data to maximize the difference between groups through analyses of variance. Thus, the k-means cluster analysis exploited the fit to small, medium, and large learning effects. Results are shown in Table 5. The k-means cluster analysis binned 41 of 135 children into the category of small effects, which was defined by back-transformed raw ES values in the range of .09-2.16 (M = 1.40; 95% CI [1.21, 1.61]). 62 of 135 children formed the category of medium effects, defined by back-transformed raw ES values in the range of 2.35-5.89 (M = 3.61; 95% CI [3.38, 3.85]). The remaining 32 children formed the category of large effects, defined by back-transformed raw ES values in the range of 6.32-27.83 (M = 10.12; 95% CI [8.79, 11.62]).

Insert Table 5 about here

Complementary ANOVAs were conducted to confirm the benchmark groups. The intent was to determine whether the benchmark groups, as defined by the k-means cluster analysis, would remain differentiated relative to other established measures of phonological learning. The measures of learning in Table 3 were again considered. ANOVAs established that the benchmark groups were statistically distinct in PCC-R pre- to posttreatment, $F(2, 132) = 11.72$, $p < .001$; in PWP pre- to posttreatment, $F(2, 132) = 4.80$, $p = .01$; and in production accuracy pre- to posttreatment, $F(2, 132) = 86.45$, $p < .001$. Similarly, the benchmark groups were statistically distinct in maintenance of learning effects, $F(2, 132) = 54.79$, $p < .001$. Thus, the boundaries

that defined small, medium, and large effects reliably differentiated generalization gain, whether indexed by ES or other independently established measures of phonological learning.

The mean ES values shown Table 5 are thus put forth as preliminary benchmarks for interpretation of single-subject research on phonological treatment. Specifically, the mean ES values of 1.4, 3.6, and 10.1 are proposed as estimates of small, medium, and large learning effects, respectively. It should be noted that the proposed benchmarks follow from mean ES values, but this does not preclude alternatives, depending on interpretation of the data at-hand. For example, it is possible to define benchmark groups using the ranges reported in Table 5 or alternatively, standard deviations from the mean following from Figure 2. It should be further emphasized that benchmark estimates herein are specific to generalization as a function of treatment and may not be applicable to other aspects of language learning. Likewise, the benchmark estimates are specific to children with phonological disorders and may not be applicable to other clinical populations.

General Discussion

This paper set out to evaluate ES for single-subject design in treatment of children with phonological disorders. The goal was to document the magnitude of generalization gain achieved by a relatively homogeneous group of children, and in doing so, to delineate preliminary boundaries and benchmarks for interpretation of ES. Toward this end, results showed ES closely aligned with other conventional measures that have been used previously to gauge children's phonological learning, either experimentally or clinically. ES further corroborated patterns of generalization that were derived from visual inspection of learning in prior experimental studies of treatment efficacy. Moreover, ES was linked to children's performance on diagnostic assessments of phonology, but not other demographic characteristics

or related linguistic skills and nonlinguistic skills. Together, the results supported ES as a valid analytic complement to single-subject design and provided initial data to derive preliminary benchmarks for interpretation of ES for the population. The results have implications for the design and interpretation of single-subject research on phonological treatment, and identify questions for future research on ES and meta-analyses of phonological treatment. These serve to frame the general discussion.

Research and Clinical Implications

The results offer some new perspectives on variability and interpretability of single-subject research on phonological treatment. On the side of variability, there are at least two observations to be made. First, the obtained range of ES and 95% CI suggest boundaries and typicality of learning from treatment. There was considerable variation in raw ES for the group (i.e., 0.9-27.83), yet the corresponding 95% CI of the mean was narrow (i.e., [3.20, 4.18]). This begins to define a possible range for the average magnitude of generalization gain that may reasonably be expected from phonological treatment. This further provides a starting point in evaluation of treatment effects from the vantage of ES.

A second observation relates to variability in baseline performance, which is relevant to computation of ES. Recall that formulas for ES uniformly rely on baseline variation to determine magnitude of gain. This presents challenges for single-subject design, which necessitates near zero-variance in the baseline to establish functional relationships between treatment and learning (Beeson & Robey, 2006; Kratochwill, 1978). Yet, in the absence of baseline variability, it is not possible to compute an ES. In the present study, baseline variability in accuracy of production of sounds excluded from the phonemic inventory was established for the study population, with the standard deviation being .02. This value was obtained from a

cohort of children representative of the population at-large. As such, it might be possible to apply the standard deviation of the baseline obtained herein to other research or to clinical practice when problematic cases of zero-variance are observed. This follows Glass (1977; Busk & Serlin, 1992) and adds to the solutions that have been offered to accommodate zero-variance in the baseline of single-subject design.

On the side of interpretation, the benchmarks that were established offer a starting point for gauging the practical significance (Bothe & Richardson, 2011) of treatment. The utility is that a given treatment or treatment condition may be described qualitatively as promoting small, medium, or large learning effects. With preliminary benchmarks in place, such characterizations might now be possible.

Preliminary benchmarks also afford for comparisons (and refinements) when applied across studies. This can be illustrated through reconsideration of ES data summarized in Table 1 relative to benchmarks reported in Table 5. Table 1 shows that preliminary applications of ES in phonological treatment ranged from 2.6-16.58. When viewed relative to the mean benchmarks in Table 5, it can be seen that the available ES data crosscut the categories of small (1.4), medium (3.6), and large (10.1) learning effects. It is possible to extend the illustration further in qualifying the efficacy of specific experimental conditions. Notice in Table 1, that three stimulus conditions were consistent with large learning effects. These included treatment of a sound in (1) in frequent words, where ES = 12.60 (Gierut & Morrisette, 2011), (2) later acquired words, where ES = 11.41-16.58 (Gierut & Morrisette, 2012a), and (3) words from dense neighborhoods comprised of many phonetically similar forms, where ES = 11.39-14.83 (Gierut & Morrisette, 2012b). Previously reported ES data, when coupled with newly established descriptive benchmarks, helps to reveal a possible set of stimulus conditions associated with greater

magnitudes of generalization gain. This illustrates how benchmarks might be used in subsequent comparisons of treatment effects within and across experimental studies.

The clinical relevance of such comparisons lies in the potential to isolate treatment conditions associated with greater magnitudes of gain. This has advantages over visual inspection of learning data because recommendations emerge from a constant scalar-free index and corresponding descriptors. This notwithstanding, it must be emphasized that ES is not meant to replace, but to complement visual inspection of learning data in single-subject design (Olive & Smith, 2005). Individual patterns of learning should be evaluated in tandem with ES to fully inform interpretation and practice. By referencing dual sources of data, the literature on evidence-based practice might be better weighed against the unique profiles and needs of individual children to thereby maximize learning in applied clinical settings.

Research and Clinical Extensions

Despite practical benefits, certain qualifications must be made. In particular, the present results bear only on ES specific to the multiple baseline design, in computation of Standard Mean Difference, relative to generalization learning, by a relatively homogeneous cohort of children, who were exposed to a uniform treatment protocol. The work was purposefully constrained to minimize extraneous variables and spurious results, but each dimension of control might be viewed conversely as a limitation to be addressed in future research.

Design and computation. The focus on the multiple baseline design was motivated by the wide use of this design in the phonological literature, suitability of this design to the Standard Mean Difference, and interpretability relative to applications of ES in other single-subject clinical research. This allowed for continuity with the broader literature, but other single-subject designs have likewise been used to determine treatment effects for phonological disorders. Of

mention are the concurrent, alternating treatments, or multiple probe designs (McReynolds & Kearns, 1983). Likewise, other nonregression and regression techniques are viable indices of ES (Campbell & Herzinger, 2010; Faith et al., 1996; Parker et al., 2007), with some formulas being better fits to specific designs and questions of interest. For example, in treatment of phonological processes (Weiner, 1981), appropriate ES formulas might include computation of percentage reduction data (Scruggs, Mastropieri & Casto, 1987) or percentage of zero data (Scotti, Evans, Meyer, & Walker, 1991). The reason is that these formulas assess the magnitude of behavior reduction, which is wholly in keeping with the suppression of phonological processes. In future research, it will be necessary to document ES for a wider range of designs that are matched to the most appropriate ES statistic. This will round out the utility of ES for single-subject design as applied to phonological treatment.

Standard Mean Difference also needs to be assessed relative to other ES formulas. The strengths and weaknesses of nonregression versus regression formulas have been discussed at length in the literature (cf. Busk & Serlin, 1992; Campbell, 2004; Faith et al., 1996; Olive & Smith, 2005; Parker et al., 2007). Yet, the different formulas have not been empirically evaluated in side-by-side comparisons of treatment for phonological or other clinical linguistic disorders. Research on developmental disability may offer a viable model for how to proceed (Campbell, 2004; Olive & Smith, 2005; see also Wolery, Busick, Reichow, & Barton, 2010 for a similar approach). In that work, a given data set was submitted to several different ES computations, with the goal of determining the insights to treatment efficacy offered by each formula. It is curious to note that the conclusions of such comparisons have been at direct odds, with one study favoring nonregression (Campbell, 2004) and another, regression (Olive & Smith, 2005) techniques. While we await comparative research for ES in treatment of phonological

disorders, it might be prudent to follow Durlak's (2009) suggestion that the choice of ES computation be guided by the purpose and methods of a given research study.

Measures of learning. The present study focused on generalization as the primary measure of phonological learning. This was motivated by the interest in inducing system-wide phonological gains as a gauge of treatment efficacy. Yet, other aspects of learning are equally relevant to the clinical process. One is a child's performance session-by-session in treatment. Session-by-session performance informs a clinician about potential modifications or adjustments to treatment that might be needed to better facilitate a child's learning. As such, the documentation of ES for session-by-session performance will be a vital complement to generalization. A related consideration in the clinical process is maintenance. Maintenance was examined for its correlation with generalization herein; however, it will be essential to establish ES independently for this aspect of learning. Future work along these lines has the potential to yield insight to the full scope of learning that is integral to the clinical process, from session-by-session performance during treatment, to generalization from treatment, and subsequently, maintenance in the absence of treatment.

Generalization was further defined relative to children's phonemic inventories with documentation of change in production of singleton sounds excluded from the repertoire. The intent was to modify the phonotactics of a child's grammar because this ensured stable baseline performance as central to the multiple baseline design. Nonetheless, it is well established that phonological disorders manifest in different ways and affect other elements of the phonological system. Children may exhibit errors in phonetic, phonemic, or syllabic structure, and phonological processes or rules may apply. Future research will need to examine children's error patterns more broadly, perhaps defining generalization relative to change in phonetic levels

of complexity (Tyler & Figurski, 1994), phonological mean length of utterance (Ingram & Ingram, 2001), sonority difference in consonant clusters (Gierut, 1999), or suppression of phonological processes (Weiner, 1981). Work of this sort has the potential to extend the applicability of ES to the multifaceted nature of the error patterns associated with phonological disorders.

Study population. Children who contributed data to the present study constituted a relatively homogeneous group in that they all met well-defined inclusionary and exclusionary criteria for participation. With exception of phonology, performance was within typical limits for other related linguistic and nonlinguistic skills. This notwithstanding, it is known that children with phonological disorders may present with co-occurring deficits, such that phonology interfaces, for example, with disfluency (Nippold, 2001) or specific language impairment (Shriberg, Tomblin, & McSweeney, 1999). Other children with phonological disorders may exhibit lags in word learning (Shriberg & Kwiatkowski, 1994) or limitations in phonological working memory (Shriberg et al., 2009). Consequently, research is needed to introduce heterogeneity into the study population by computing ES for children with varying skills. While heterogeneity might broaden representation, it carries the risk of introducing noise in the data that may cloud ES results.

Homogeneity of the study population may have further impacted the detection of contributing variables. Recall that ES was associated only with diagnostic tests of phonological skills and change in measures of phonological learning. While it is possible that ES reflects only phonological factors, this must be confirmed by exploring additional variables as possible contributors to ES. For example, stimulability (Powell et al., 1998) and imitation (Dean et al., 1995) might be considered because of reported influences on phonological learning.

Metalinguistic skills might also be considered due to close connections with phonological learning, specifically (Rvachew & Grawburg, 2006) and language learning, generally (Carroll, Snowling, Hulme, & Stevenson, 2003). To home in on such contributing factors, it will be necessary to broaden the baseline assessments that are used to identify the population in future research. If contributing variables were to be identified, treatment might be designed to take further advantage of child-specific skills to boost phonological learning.

It may also be relevant to consider phonological learning in typical development as a potential platform for comparison. Children with typical development might be followed longitudinally to document phonological advances that take place naturalistically. Consideration might be given to age- or phonological-matching of participants. Likewise, accuracy of production might be traced for incorrect and also correct sounds relative to advances in lexical size and/or grammatical complexity. This is because phonological selection/avoidance, segmental trade-offs, progressive idioms, and other creative strategies are often observed in typical development (Vihman et al., 1986) as a child's lexical representations become more segmentally analyzable. ES for typical development may reveal the magnitude of phonological gain and the variance expected in the absence of a disorder and treatment. From this, new information may emerge about the process of phonological learning in typical versus atypical development.

Treatment protocol. The treatment protocol was another constant of the present study. This ensured that children received the same duration of treatment using the same format of instruction. Recall that treatment was capped at 19 sessions (hours), whereas treatment delivery in clinical settings is likely to continue until a child normalizes, independent of the time course. Recall too that treatment herein centered on imitation and spontaneous production of sounds.

While consistent with conventional procedures, there are a variety of instructional techniques available for phonological treatment (Baker & McCloud, 2011; Brumbaugh & Smit, 2013; Williams, McLeod, & McCauley, 2010 for reviews). Examples include treatment directed toward perception, stimulability, formation of phonological categories, or phonological awareness. Treatment by caregivers or in group settings are other options for service delivery. Continued research will need to establish ES for these various treatment options. Work of this sort might help to identify the instructional techniques that are most beneficial to children in phonological treatment. This would have further consequences for understanding the effectiveness and efficiency of phonological treatment as necessary complements to the evaluation of treatment effects herein.

It is clear that research opportunities in the study of ES in single-subject designs for phonological treatment are abundant. As ES data accrue, a foundation will be set for meta-analyses of single-subject designs in evaluation of the efficacy of phonological treatment. Such data will bring us closer to pinpointing the extent to which treatment promotes phonological learning, the child-specific and external factors that impact that learning, and the ways that treatment might be administered to maximize the learning process. These are fundamental issues at the core of evidence-based practice.

References

Bain, B. A., & Dollaghan, C. A. (1991). The notion of clinically significant change. *Language, Speech and Hearing Services in Schools*, *22*, 264-270.

Baker, E., & McLeod, S. (2011). Evidence-based practice for children with speech sound disorders: Part 1 narrative review. *Language, Speech and Hearing Services in Schools*, *42*, 102-139.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *24*, 215-220.

Beeson, P., & Robey, R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review*, *16*, 161-169.

Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *American Journal of Speech-Language Pathology*, *20*, 233-242.

Brumbaugh, K., & Smit, A. (2013). Treating children ages 3-6 who have speech sound disorder: A survey. *Language, Speech and Hearing Services in Schools*, *44*, 306-319.

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* (pp. 187-212). NJ: Lawrence Erlbaum.

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology*, *21*, 397-414.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification*, *28*, 234-246.

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 417-453). New York: Routledge.

Carroll, J. M., Snowling, M. J., Stevenson, J., & Hulme, C. (2003). The development of phonological awareness in preschool children. *Developmental Psychology*, *39*, 913.

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

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment*, *9*, 141-150.

Dean, E. C., Howell, J., Waters, D., & Reid, J. (1995). Metaphon: A metalinguistic approach to the treatment of phonological disorder in children. *Clinical Linguistics & Phonetics*, *9*, 1-19.

Dinnsen, D. A. (1984). Methods and empirical issues in analyzing functional misarticulation. In M. Elbert, D. A. Dinnsen & G. Weismer (Eds.), *Phonological theory and the misarticulating child (ASHA Monographs No. 22)* (pp. 5-17). Rockville, MD: American Speech-Language-Hearing Association.

Dollaghan, C. A. (2007). *Handbook for evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*, 1136-1146.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test–revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines,

MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test* (4th ed.). Circle Pines, MN: American Guidance Service.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*, 917-928.

Edeal, D. M., & Gildersleeve-Neumann, C. E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *American Journal of Speech-Language Pathology*, *20*, 95-110.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B., Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Hillsdale, NJ: Lawrence Erlbaum.

Gast, D. L. (Ed.) (2010). *Single subject research methodology in behavioral sciences*. New York: Routledge.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In Gast, D. L. (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199-233). New York: Routledge.

Gierut, J. A. (1998). Treatment efficacy: Functional phonological disorders in children. *Journal of Speech, Language, and Hearing Research*, *41*, S85-S100.

Gierut, J. A. (1999). Syllable onsets: Clusters and adjuncts in acquisition. *Journal of Speech, Language, and Hearing Research*, *42*, 708-726.

Gierut, J. A. (2001). Complexity in phonological treatment: Clinical factors. *Language, Speech and Hearing Services in Schools*, *32*, 229-241.

Gierut, J. A. (2007). Phonological complexity and language learnability. *American Journal of*

Speech-Language Pathology, *16*, 6-17.

Gierut, J. A. (2008a). Fundamentals of experimental design and treatment. In D. A. Dinnsen & J. A. Gierut (Eds.), *Optimality theory, phonological acquisition and disorders* (pp. 93-118). London: Equinox.

Gierut, J. A. (2008b). Phonological disorders and the developmental phonology archive. In D. A. Dinnsen & J. A. Gierut (Eds.), *Optimality theory, phonological acquisition and disorders* (pp. 37-92). London: Equinox.

Gierut, J. A., & Morrisette, M. L. (2011). Effect size in clinical phonology. *Clinical Linguistics & Phonetics*, *25*, 975-980.

Gierut, J. A., & Morrisette, M. L. (2012a). Age-of-word acquisition effects in treatment of children with phonological delays. *Applied Psycholinguistics*, *33*, 121-144.

Gierut, J. A., & Morrisette, M. L. (2012b). Density, frequency and the expressive phonology of children with phonological delay. *Journal of Child Language*, *39*, 804-834.

Gierut, J. A., & Morrisette, M. L. (2014). How to meet the neighbors: Modality effects on phonological generalization. *Clinical Linguistics & Phonetics*, *28*, 477-492.

Gierut, J. A., Morrisette, M. L., & Ziemer, S. (2010). Nonwords and generalization in children with phonological disorders. *American Journal of Speech-Language Pathology*, *19*, 167-177.

Gierut, J. A., Simmerman, C. L., & Neumann, H. J. (1994). Phonemic structures of delayed phonological systems. *Journal of Child Language*, *21*, 291-316.

Glass, G. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*, 351-379.

Goldman, R., & Fristoe, M. (1986). *Goldman-Fristoe test of articulation*. Circles Pines, MN:

American Guidance Service.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe test of articulation* (2nd ed.). Circle Pines, MN: American Guidance Service.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press.

Hesketh, A., Adams, C., Nightingale, C., & Hall, R. (2000). Phonological awareness therapy and articulatory training approaches for children with phonological disorders: A comparative outcome study. *International Journal of Language & Communication Disorders*, *35*, 337-354.

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

Hoyle, R. H. (Ed.). (1999). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1981). *Test of early language development*. Austin, TX: Pro-Ed.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1991). *Test of early language development* (2nd ed.). Austin, TX: Pro-Ed.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). *Test of early language development* (3rd ed.). Austin, TX: Pro-Ed.

Ingram, D. (2002). The measurement of whole word productions. *Journal of Child Language*, *29*, 713-734.

Ingram, D., & Ingram, K. D. (2001). A whole-word approach to phonological analysis and

intervention. *Language, Speech and Hearing Services in Schools*, *32*, 271-283.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.

Kirk, S. A., McCarthy, J. J., & Kirk, W. D. (1968). *Illinois test of psycholinguistic abilities–revised*. Chicago: University of Illinois Press.

Komrey, J., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education*, *65*, 73-93.

Kratchowill, T. R. (1978). *Single subject research: Strategies for evaluating change*. New York: Academic Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33,* 159-74.

Law, J., Garrett, Z., & Nye, C. (2004). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *47*, 924-943.

Levine, M. N. (1986). *Leiter international performance scale.* Chicago: Stoelting.

Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology*, *13*: 119. Retrieved from http://www.biomedcentral.com/1471-2288/13/119

McReynolds, L. V., & Kearns, K. P. (1983). *Single-subject experimental designs in communicative disorders*. Baltimore, MD: University Park Press.

Miccio, A. W., & Elbert, M. (1996). Enhancing stimulability: A treatment program. *Journal of*

Communication Disorders, *29*, 335-351.

Miccio, A. W., Gallagher, E., Grossman, C. B., Yont, K. M., & Vernon-Feagans, L. (2001). Influence of chronic otitis media on phonological acquisition. *Clinical Linguistics & Phonetics*, *15*, 47-51.

Morrisette, M. L., & Gierut, J. A. (2002). Lexical organization and phonological change in treatment. *Journal of Speech, Language, and Hearing Research*, *45*, 143-159.

Morrisette, M. L., Hoover, J. R., & Gierut, J. A. (2012, June). *Lexical neighborhoods in recognition by children with phonological disorders*. 33rd Annual Symposium on Research in Child Language Disorders, Madison, WI.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*, 622-637.

Newcomer, P. L., & Hammill, D. D. (1988). *Test of language development–primary* (2nd ed.). Austin, TX: Pro-Ed.

Newcomer, P. L., & Hammill, D. D. (1997). *Test of language development–primary* (3rd ed.). Austin, TX: Pro-Ed.

Nippold, M. A. (2001). Phonological disorders and stuttering in children: What is the frequency of co-occurrence? *Clinical Linguistics & Phonetics*, *15*, 219-228.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology*, *25*, 313-324.

Olswang, L. B. (1998). Treatment efficacy research. In C. M. Frattali (Ed.), *Measuring outcomes in speech-language pathology* (pp. 134-150). New York: Thieme.

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

Powell, T. W. (1991). Planning for phonological generalization: An approach to treatment target selection. *American Journal of Speech-Language Pathology*, *1*, 21-27.

Powell, T. W., Elbert, M., Miccio, A. W., Strike-Roussos, C., & Brasseur, J. (1998). Facilitating [s] production in young children: An experimental evaluation of motoric and conceptual treatment approaches. *Clinical Linguistics & Phonetics*, *12*, 127-146.

Robbins, J., & Klee, T. (1987). Clinical assessment of oropharyngeal motor development in young children. *Journal of Speech and Hearing Disorders*, *52*, 271-277.

Robey, R. R. (1994). The efficacy of treatment for aphasic persons: A meta-analysis. *Brain and Language*, *47*, 582-608.

Robey, R. R. (1998). A meta-analysis of clinical outcomes in the treatment of aphasia. *Journal of Speech, Language, and Hearing Research*, *41*, 172-187.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Review: Single-subject clinical-outcome research: designs, data, effect sizes, and analyses. *Aphasiology*, *13*, 445-473.

Roid, G. H., & Miller, L. J. (1997). *Leiter international performance scale–revised*. Chicago: Stoelting.

Rosenthal, R., & Rosnow, R. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). Boston: Mc-Graw Hill.

Rvachew, S., & Grawburg, M. (2006). Correlates of phonological awareness in preschoolers with speech sound disorders. *Journal of Speech, Language, and Hearing Research*, *49*, 74-87.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American*

Journal on Mental Retardation, *93*, 233-256.

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

Shriberg, L. D. & Kwiatkowski, J. (1994). Developmental phonological disorders I: A clinical profile. *Journal of Speech and Hearing Research*, *37*, 1100-1126.

Shriberg, L. D., & Lof, G. L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics & Phonetics*, *5*, 225-279.

Shriberg, L. D., Kwiatkowski, J., & Gruber, F. A. (1994). Developmental phonological disorders II: Short-term speech-sound normalization. *Journal of Speech and Hearing Research*, *37*, 1127-1150.

Shriberg, L. D., Lohmeier, H. L., Campbell, T. F., Dollaghan, C. A., Green, J. R., & Moore, C. A. (2009). A nonword repetition task for speakers with misarticulations: The syllable repetition task (SRT). *Journal of Speech, Language, and Hearing Research*, *52*, 1189-1212.

Shriberg, L. D., Tomblin, J. B., & McSweeny, J. L. (1999). Prevalence of speech delay in 6-year-old children and comorbidity with language impairment. *Journal of Speech, Language, and Hearing Research*, *42*, 1461-1481.

Shriberg, L., Austin, D., Lewis, B., McSweeny, J., & Wilson, D. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research*, *40*, 708-722.

Stoel-Gammon, C. (1985). Phonetic inventories, 15-24 months: A longitudinal study. *Journal of Speech and Hearing Research*, *28*, 505-512.

Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject–design intervention

research for students with LD. *Journal of Learning Disabilities*, *33*, 114-136.

Thompson, C. K., den Ouden, D. B., Bonakdarpour, B., Garibaldi, K., & Parrish, T. B. (2010). Neural plasticity and treatment-induced recovery of sentence processing in agrammatism. *Neuropsychologia*, *48*, 3211-3227.

Tyler, A. A., & Figurski, G. R. (1994). Phonetic inventory changes after treating distinctions along an implicational hierarchy. *Clinical Linguistics & Phonetics*, *8*, 91-108.

Vihman, M. M., Ferguson, C. A., & Elbert, M. (1986). Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, *7*, 3-40.

Weiner, F. F. (1981). Treatment of phonological disability using the method of meaningful minimal contrast: Two case studies. *Journal of Speech and Hearing Disorders*, *46*, 97-103.

Wiig, E. H., Secord, W., & Semel, E. (1992). *Clinical evaluation of language fundamentals–preschool*. San Antonio, TX: Harcourt Brace.

Wiig, E. H., Secord, W., & Semel, E. (2004). *Clinical evaluation of language fundamentals–preschool* (2nd ed.). San Antonio, TX: Harcourt Brace.

Williams, A. L., McLeod, S., & McCauley, R. J. (Eds.). (2010). *Interventions for speech sound disorders in children*. Baltimore: Paul H. Brookes.

Williams, K. T. (1997). *Expressive vocabulary test*. Circle Pines, MN: American Guidance Service.

Williams, K. T. (2007). *Expressive vocabulary test* (2nd ed.). Circle Pines, MN: American Guidance Service.

Wolery, M. Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison for overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*, 18-28.

Table 1

Multiple Baseline Studies of Phonological Treatment Reporting ES.

Study & Sample Size	Stimulus Conditions	ES
Gierut & Morrisette (2011)	Frequent words	12.6
N = 8	Infrequent words	5.9
	Dense words	2.6
	Sparse words	4.3
Gierut & Morrisette (2012a)	Early acquired-frequent words	2.81
N = 10	Early acquired-infrequent words	3.66
	Late acquired-frequent words	11.41
	Late acquired-infrequent words	16.58
Gierut & Morrisette (2012b)	Dense-frequent words	14.83
N = 8	Dense-infrequent words	11.39
	Sparse-frequent words	3.19
	Sparse-infrequent words	5.31

Gierut & Morrisette (2014)	Auditory-visual input	7.92
N = 9	Auditory input	7.04
	Visual input	2.77

Table 2

Conventional Measures of Phonological Learning Relative to ES.

Dependent Measure	Range	*M* [95% CI]	Statistic
Percent Consonants Correct-Revised Pre-Post[a, b]	–.16-.33	.05 [.04, .06]	$r(133) = .40, p < .001$
Proportion of Whole Word Proximity Pre-Post[b, c]	–.15-.15	.03 [.02, .04]	$r(133) = .29, p = .001$
Production Accuracy Pre-Post[b, d]	–.05-.81	.09 [.08, .11]	$r(133) = .81, p < .001$
Maintenance[d]	.00-1.00	.16 [.14, .18]	$r(133) = .76, p < .001$

NOTE: CI = confidence interval. Range, *M*, and 95% CI are proportional values.

[a]Shriberg et al., 1997

[b]Values for the range, *M*, and 95% CI represent difference scores.

[c]Ingram & Ingram, 2001

[d]Data were log transformed to approximate a normal distribution; back-transformed values are reported for the range, *M,* and 95% CI.

Table 3

Summary of Potential Variables Contributing to ES.

Contributing Variables	*N*[a]	Range	*M* [95% CI]	Statistic
Diagnostic Assessments of Phonology				
Goldman-Fristoe (percentile)[b]	135	–1-16	3.52 [2.85, 4.19]	$r(133) = .21$, $p = .02$*
Percent Consonants Correct-Revised[c]	135	.10-.78	.49 [.46, .51]	$r(133) = .33$, $p < .001$*
Proportion of Whole Word Proximity[d]	135	.45-.91	.73 [.71, .74]	$r(133) = .33$, $p < .001$*
Phonemic inventory size	135	7-20	14.25 [13.75, 14.75]	$r(133) = .27$, $p = .001$*
Demographics				
Age (months)	135	36-93	52.99 [51.29, 54.69]	$r(133) = .09$, $p = .32$
Sex	89 M/46 F			$t(133) = –.15$, $p = .88$
Onset of first words (typical/delayed)[e]	95/12			$t(105) = .49$, $p =. 63$
Family history (yes/no)[e]	62/63			$t(123) = –.13$, $p = .90$
Otitis media history (yes/no)[e]	54/75			$t(127) = .63$, $p = .53$

Diagnostic Assessments of Related Skills				
Oral-motor function[f]	135	93-112	108.78 [108.13, 109.43]	$r(133) = .02$, $p = .82$
Receptive vocabulary[g]	135	72-138	107.92 [105.75, 110.09]	$r(133) = –.01$, $p = .92$
Expressive vocabulary[h]	40	94-126	108.70 [105.59, 111.81]	$r(38) = –.29$, $p = .07$
Receptive/expressive language[i]	135	78-137	110.76 [108.45, 113.08]	$r(133) = .04$, $p = .69$
Nonverbal intelligence[j, k]	135	82-169	121.92 [118.99, 124.85]	$r(133) = –.00$, $p = .99$
Cognitive-social rating[k]	40	70-117	97.83 [94.53, 101.12]	$r(38) = –.10$, $p = .55$
Forward digit span[l]	48	27-53	36.35 [34.50, 38.21]	$r(46) = .01$, $p = .93$
Nonword repetition[m]	29	32-77	54.38 [49.97, 58.79]	$r(27) = .06$, $p = .74$
Memory screen[k]	40	81-143	107.05 [102.08, 112.02]	$r(38) = –.17$, $p = .30$

Note: CI = confidence interval.

[a]The diagnostic battery was updated over the 30-year duration of the research program to reflect current test editions and standards of assessment for phonological disorders. This explains why sample sizes may have differed across diagnostic measures; nonetheless, a common core set of diagnostic results was available for all children.

[b]Goldman & Fristoe, 1986, 2000; percentile scores are reported.

[c]Shriberg et al., 1997; range, *M*, and 95% CI are proportional values.

[d]Ingram & Ingram, 2001; range, *M*, and 95% CI are proportional values.

[e]Not all parents chose to report this information on the child history questionnaire.

[f]Clinical assessment of oropharyngeal motor development in young children (Robbins & Klee, 1987)

[g]*Peabody picture vocabulary test* (Dunn & Dunn, 1981, 1997, 2007)

[h]*Expressive vocabulary test* (Williams, 1997, 2007)

[i]*Clinical evaluation of language fundamentals–preschool* (Wiig, Secord, Semel, 1992, 2004); *Test of early language development* (Hresko, Reid, & Hammill, 1981, 1991, 1999); *Test of language development–primary* (Newcomer & Hammill, 1988, 1997)

[j]*Leiter international performance scale* (Levine, 1986)

[k]*Leiter international performance scale–revised* (Roid & Miller, 1997)

[l]*Illinois test of psycholinguistic abilities–revised* (Kirk, McCarthy, & Kirk, 1968)

[m]Nonword repetition task (Dollaghan & Campbell, 1998)

Table 4

Time and Performance in Treatment Relative to ES.

	Range	*M* [95% CI]	Statistic
Total sessions	5-19	13.47 [12.64, 14.30]	$r($133) = –.03, p = .76
Sessions in imitation phase	2-7	5.36 [5.03, 5.69]	r(133) = .00, p = .99
Sessions in spontaneous phase	3-12	8.11 [7.51, 8.71]	r(133) = –.04, p = .66
Max % accuracy in imitation phase	24-100	82.18 [79.88, 84.48]	r(133) = .09, p = .29
Max % accuracy in spontaneous phase	56-100	93.87 [92.59, 95.14]	r(133) = –.10, p = .25
Number of probes[a]	4-16	10.01 [9.47, 10.54]	r(133) = –.11, p = .22

Note: CI = confidence interval.

[a]Probes administered in baseline and during treatment only.

Table 5

Estimated Small, Medium, and Large Learning Effects in Single-Subject Design for Phonological Treatment.

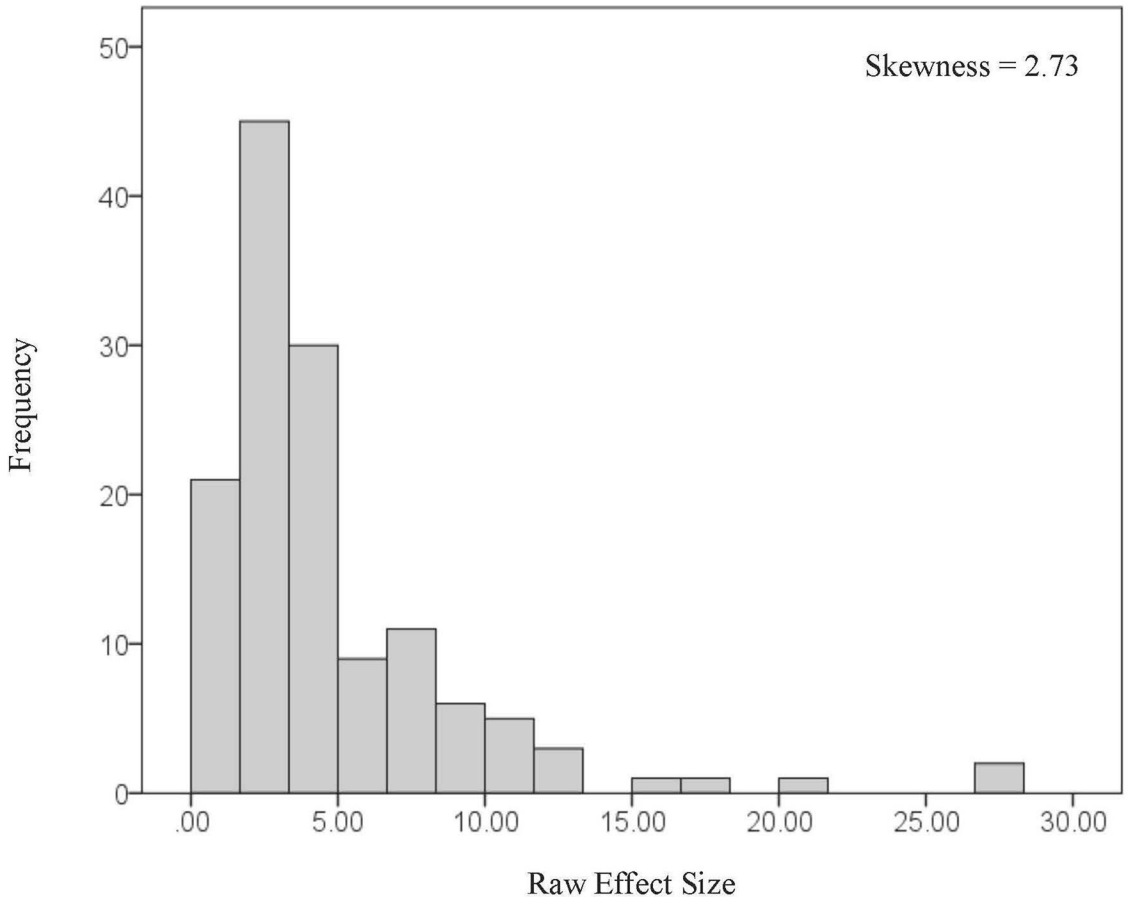
	n	Range	*M* [95% CI]
Small	41	.09-2.16	1.40 [1.21, 1.61]
Medium	62	2.35-5.89	3.61 [3.38, 3.85]
Large	32	6.32-27.83	10.12 [8.79, 11.62]

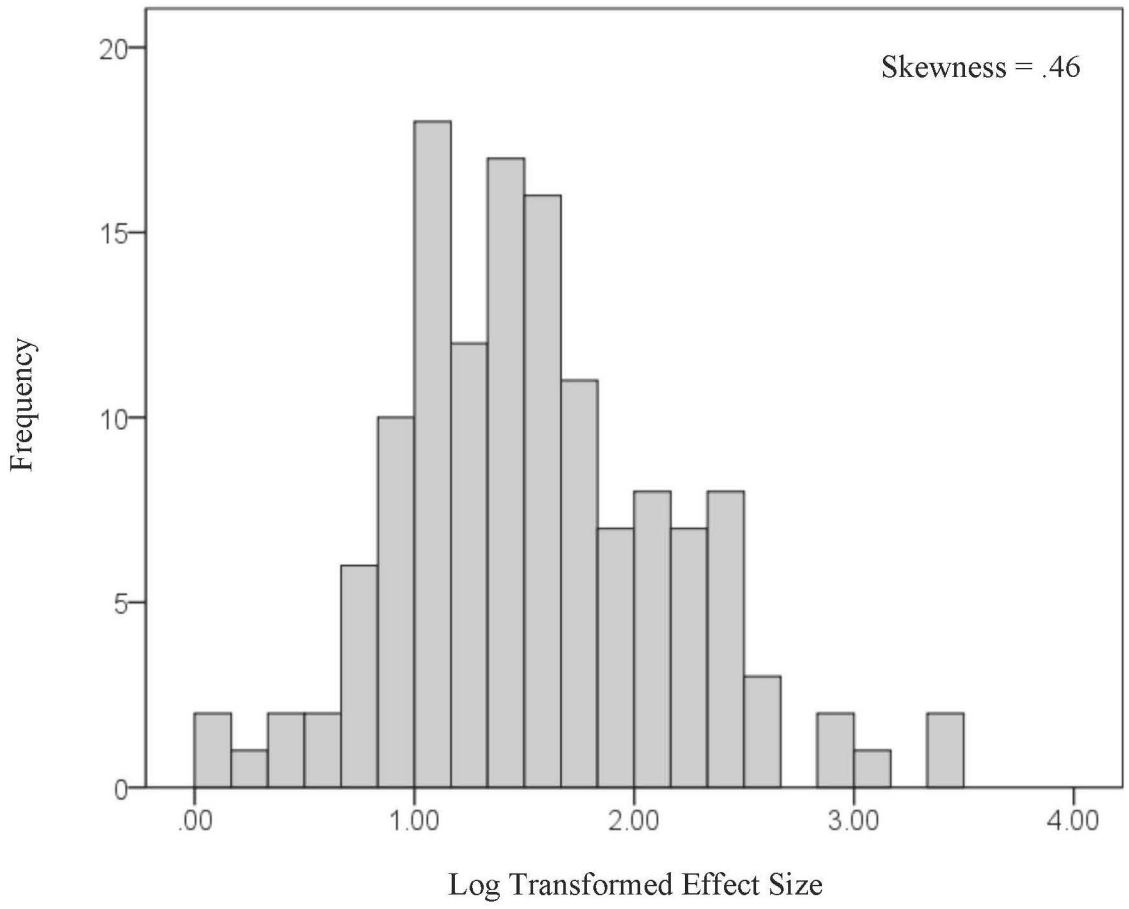
Note: CI = confidence interval.

Figure Captions

Figure 1. Histogram of raw ES for the study population.

Figure 2. Histogram of log transformed ES for the study population. For interpretation, log values were back-transformed to establish the mean raw ES for children with phonological disorders as 3.66, 95% CI [3.19, 4.18].


Skewness = 2.73
Frequency
50
40
30
20
10
0
.00
5.00
10.00
15.00
20.00
25.00
30.00
Raw Effect Size


Skewness = .46
Frequency
20
15
10
5
0
.00
1.00
2.00
3.00
4.00
Log Transformed Effect Size

Appendix

Treatment Protocol

1. Administer baseline probe for 2+ sessions based on order of enrollment
2. Treatment begins
    a. Production training
        i. Experimenter shows Child picture of production stimulus
        ii. Experimenter models production
        iii. Child responds in imitation
        iv. Experimenter provides feedback about accuracy
        v. Next trial is initiated; repeat b. i-v
        vi. Continue for 1-hr, approximately 90 trials
    b. Administer probe as required following a ratio schedule averaging every third session
    c. Child dismissed until next session
3. Repeat 2 above for 7 total sessions or until Child achieves 75% accuracy of production of treated stimuli over 2 consecutive sessions, whichever occurs first
4. After completion of 3, administer probe in the very next session
5. Repeat 2 above in all subsequent sessions, but require Child to spontaneously produce production stimuli without Experimenter model (i.e., omit a. ii above)
    a. Continue for 12 total sessions or until Child achieves 90% accuracy of production of treated stimuli over 3 consecutive sessions, whichever occurs first
6. After completion of 5, administer probe in the very next session
7. Treatment is completed; experiment ends
8. Re-administer probe after treatment is withdrawn, continuing approximately 8 weeks