Published in final edited form as:
Clin Linguist Phon. 2011 November ; 25(0): 975–980. doi:10.3109/02699206.2011.601392.

Effect size in clinical phonology

Judith A. Gierut and **Michele L. Morrisette**
Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN, USA

Abstract

The purpose of this article is to motivate the use of effect size (ES) for single-subject research in clinical phonology, with an eye towards meta-analyses of treatment effects for children with phonological disorders. Standard mean difference (SMD) is introduced and illustrated as one ES well suited to the multiple baseline (MBL) design and evaluation of generalization learning, both of which are key to experimental studies in clinical phonology.

Keywords

phonological treatment; generalization; evidence-based practice; meta-analysis

Introduction

Clinical phonology has provided an important venue for the empirical evaluation of linguistic and psycholinguistic theories, with applied consequences for the diagnosis and treatment of a range of populations with language disorders. In the study of children with phonological disorders in particular, single-subject design has been a staple of prospective experimental research because it allows for manipulation of linguistic or psycholinguistic properties in treatment as the independent variable, with evaluation of children's generalization learning as the dependent variable. While this work has contributed in practical ways to the efficacy of clinical treatment, the inherent nature of single-subject design has precluded meta-analyses of treatment effects. Consequently, it has not been possible to directly compare the magnitude of treatment effects within or across children, experimental conditions or studies. Innovations in single-subject research have introduced effect size (ES) for small-*n* studies to enable precisely these sorts of cross-comparisons. ES is a reference-free statistic that reflects the degree of change from a null (baseline) state. To our knowledge, ES has not been widely applied in studies of clinical phonology. The purpose of this article is to describe and illustrate *standard mean difference* (SMD; Busk and Serlin, 1992) as one ES computation for clinical phonology. We begin with an overview of experimental applications of single-subject design in treatment of phonological disorders to motivate the discussion of SMD. We then describe its statistic *d*, apply it to published generalization data (Morrisette and Gierut, 2002) and offer directions for future applications in clinical phonology.

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

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Treatment Designs in Clinical Phonology

Four levels of evidence have been described in evaluation of the efficacy of experimental clinical research (ASHA, 2004). At the extremes, the most robust data come from meta-analyses of randomized controlled trials (RCTs) and the least from clinical experience. In clinical phonology, only a small set of studies meets the definition of a RCT (Law, Garrett, and Nye, 2004). Instead, the main evidence to support the efficacy of phonological treatment comes from single-subject experiments. Under the scheme of evidence-based practice, single-subject studies are defined as *quasi-experimental*. This is a technical term that means the participants are not randomly assigned to experimental conditions (Campbell and Stanley, 1963). It does not imply that single-subject studies fail to meet the rigour or necessary and sufficient conditions of an experiment.

Of single-subject design options, the multiple baseline (MBL) is commonly used in clinical phonology, with replications across participants or behaviours. The design has been utilized in tests of different methods of treatment, for example, traditional or metaphon instruction (Powell, Elbert, Miccio, Strike-Roussos, and Brasseur, 1998). It has also been utilized in tests of basic linguistic and psycholinguistic constructs, for example, underlying representations (Gierut, Elbert, and Dinnsen, 1987) or lexical/sub-lexical properties of words (Morrisette and Gierut, 2002).

The assumptions and set-up of the MBL have been described in detail elsewhere (McReynolds and Kearns, 1983), but two aspects of the design warrant consideration given their relevance to ES in the context of clinical phonology. A first point is that the MBL requires stability of baseline performance to establish causal relationships between the independent and dependent variables, and to rule out maturation as an extraneous influence. Typically, studies in clinical phonology report baseline performance at near 0% production accuracy (e.g. Powell et al., 1998), with minimal fluctuation not to exceed ±10% (McReynolds and Kearns, 1983). This requirement presents a challenge because most ES formulas rely on standard deviations (*SD*s), and it is not possible to compute *SD*s on data with no variance.

A second point is that, in MBL studies of clinical phonology, treatment is the means to experimentally induce change, session-by-session, in production accuracy of the treated sound in treated word positions in treated stimuli. The primary data of interest, however, are not typically associated with session-by-session learning. Instead, the success of phonological treatment is gauged by generalization, with emphasis on system-wide improvements that promote production accuracy and increased size of the phonemic inventory (e.g. Gierut et al., 1987). Generalization is often measured by sampling treated and untreated erred sounds using a structured probe that is administered longitudinally throughout treatment, but is independent of treatment itself. This focus on generalization narrows the available ES options because many formulas evaluate session-by-session performance with an interest in assessing the slope or rate of learning (e.g. percentage of non-overlapping data, Scruggs, Mastropieri, and Casto, 1987) in lieu of overall generalization gain. SMD is one ES that has provisions for both 0%baseline data and evaluation of generalization effects, and as such, it is a suitable match to MBL applications in clinical phonology.

SMD Applied to Clinical Phonology

Debates about the advantages and disadvantages of regression versus non-regression ESs in single-subject research have led to a consensus that non-regression techniques are more conservative and tend not to overestimate treatment effects (Busk and Serlin, 1992; Campbell, 2004; Olive and Smith, 2005). Regression techniques, on the other hand, are

based on assumptions of normality, equal variance and serial independence of data, all of which are violated in single-subject designs. Among non-regression techniques, SMD has emerged as optimal (Olive and Smith, 2005) because it utilizes the richest set of data, is easy to compute, is readily interpretable and leads to a most conservative estimate of gain.

SMD (Busk and Serlin, 1992) is typically computed in the following way: the mean of baseline data (M_A) is calculated, along with the mean of generalization data collected longitudinally over the duration of treatment (M_B). The difference in means forms the numerator of the operation. The denominator is the *SD* of the baseline (SD_A), which when divided yields *d* as the ES value. *d* is computed for each leg of the MBL (i.e. participant, behaviour), and these values may then be averaged to arrive at a mean ES for each experimental condition. There are two variants of SMD: one utilizes a limited data set consisting of the first baseline and last three points of generalization data, whereas the other incorporates all baseline and generalization data into the computation. The latter is preferred because it maximizes the data as presented herein.

Given the formula, it is apparent that 0% baseline performance could be problematic as was noted. One recommendation is that baseline data be pooled across all participants of an experiment to derive the *SD* as the denominator ($SD_{\text{A–pooled across Ss}}$; S. Dickinson, Personal Communication, 3 May 2010) as in Equation (1). This follows from the assumption that participants of a given experiment form a relatively homogeneous group because all met precise inclusionary and exclusionary criteria. When the *SD* is computed on pooled baseline data, it is an actual reflection of baseline variability for the experimental population. This is preferred to a fixed (dummy) integer as an artificial estimate of baseline variability. It is also preferred to pooling baseline and treatment data (Beeson and Robey, 2006), which tends to inflate variability, given that the goal of the MBL is to accelerate performance over baseline. In all, pooled baseline data circumvent potential problematic cases of 0% performance by capturing actual variability within and across children and conditions of an experiment. Moreover, this serves as a correction for continuity, if applied uniformly in computations of SMD:

$$d = \frac{(M_B - M_A)}{SD_{\text{A–pooled across Ss}}} \quad (1)$$

Data fromMorrisette and Gierut (2002) are used to illustrate the computation of *d* in clinical phonology using the formula in Equation (1). In that study, eight preschoolers with phonological disorders were each treated on one (erred) sound in the initial position of 10 stimulus words. Two properties of the stimuli, word frequency and neighbourhood density, were manipulated as independent variables. Frequency reflects how often a given word occurs in the language, and density estimates the number of phonetically similar counterparts to a given word. Both have been linked to gains in expressive phonology (Stoel-Gammon, 2011). In Morrisette and Gierut (2002), four children were assigned to treatment of high-frequency words versus low-frequency words, and four others to words from dense neighbourhoods versus sparse neighbourhoods. A main finding was that treatment of high-frequency words and words from sparse neighbourhoods led to greater phonological generalization. This was based on visual inspection of descriptive data associated with children's generalization learning.

Table I shows the data from Morrisette and Gierut (2002) that were entered into the computation of *d*. The data herein differ somewhat from the original to provide a straightforward illustration and to comply with the formula for *d* (e.g. generalization was not subdivided into treated, within- and across-class change, data were inclusive of all baseline

and probe samples). Table I reports average percent accuracy baseline and generalization data for each child, the *SD* of pooled baseline data in correction for continuity, as well as *d* values for each child and each experimental condition. For illustration purposes, we consider data from Child 1 in computation of *d*. This child's mean percent accuracy during baseline was 1.212 and during generalization, 27.241; the difference between means is 26.029. When divided by 1.620, which is the *SD* of pooled baseline data, the ES for Child 1 is 16.07. Computation of *d* for the other children in Table I was achieved in the same way.

Once the ES is in hand, its intended use and interpretation is three-fold. First, an ES statistically confirms descriptive results of single-subject studies developed from visual inspection. Second, an ES extends descriptive results by informing the relative magnitude of gain obtained under each experimental condition. Third, when compared against an empirically generated standard, an ES reveals the strength of the treatment effects, deemed small, medium or large, for a given population.

Returning to the Morrisette and Gierut (2002) illustration, two of three interpretations may be gleaned from Table I. Specifically, the ESs are consistent with the published report, identifying high-frequency words and words from sparse neighbourhoods as optimal to generalization. This can be seen in greater *d* values for Children 1 and 2 treated on high-frequency words, relative to Children 3 and 4 treated on low-frequency words. Similarly, Children 7 and 8 had greater *d* values than Children 5 and 6 treated on words from sparse neighbourhoods versus dense neighbourhoods, respectively. Further, the ESs expand the published report when the *d* values of the experimental condition are compared. It can be seen that the magnitude of generalization gain derived from high-frequency words was twice that of low-frequency words ($d = 12.6$ vs. 5.9, respectively). A near two-fold magnitude of gain was also obtained between low- and high-density neighbourhoods ($d = 4.3$ vs. 2.6, respectively). Of greater interest is the comparison of ESs across conditions. Notice, for example, that the magnitude of gain from treatment of high-frequency words was nearly five times that of dense neighbourhoods ($d = 12.6$ vs. 2.6, respectively). Such comparisons quantify the relative magnitude of generalization gain within and across experimental conditions, thereby yielding a transparent assessment of treatment efficacy. From Table I, it is not possible to code the strength of treatment effects as small, medium or large because standards for interpretation of ES in clinical phonology have not yet been established. This, then, brings us to future research needs.

Future Directions and Cautions

ESs generally and *d* in particular are scale-free indices with no corresponding probability values. This is attractive because it allows for direct comparisons of ES across studies for purposes of meta-analyses; however, it is also a drawback because benchmarks for interpretation are necessitated. While benchmarks are available for between-group designs (Cohen, 1988), it is inappropriate to extend these to small-*n* studies. Likewise, while ES standards are emerging for specific disorders (Beeson and Robey, 2006), they are not necessarily generalizable across populations (Durlak, 2009). Thus, for clinical phonology, a standard against which to evaluate ES must be established empirically, defined by confidence intervals. This can be achieved only by generating ES data from multiple children and studies. To our knowledge, there is just one single-subject study on phonological disorders that documents ES (Gierut and Morrisette, 2011). As a first step, one suggestion is that single-subject studies of phonological treatment report ESs as conventional. As reports accumulate, it should be possible to identify a range of ESs, across studies and for the population, from which benchmarks may be delineated.

Despite potential advantages in the report of ES in single-subject research, there are three cautionary points. First, different ES computations are available, each with its own set of assumptions. These may be more or less favourably matched to specific experimental designs and research questions. SMD was introduced herein, but the choice of ES should be guided by the specific questions of interest. As such, ES techniques may vary across studies even within a given population. Second, applications of ES in single-subject design should not come at the expense of conventional interpretations that rely on visual inspection of learning data. These should remain intact, complemented by *d* or other ES statistic. Finally, ES may be taken at face value when differentially evaluating experimental variables associated with theoretical questions in clinical phonology. However, in practice, ES might not fully capture the value of treatment for the overall well-being of a given child. Because experimental research on clinical phonology informs evidence-based practice, the theoretical and clinical significance of treatment effects must be weighed and interpreted in tandem.

Acknowledgments

We applaud Martin Ball for his encouragement of interdisciplinary theoretical study of clinical populations and his reception to innovations that break new ground in understanding language structure, acquisition and breakdown.

This research was supported in part by a grant from the National Institutes of Health (DC001694) awarded to Indiana University.

References

American Speech-Language-Hearing Association. [Accessed 1 June 2011] Evidence-based practice in communication disorders: An introduction (Technical Report). 2004. from: www.asha.org/policy/

Beeson P, Robey R. Evaluating single-subject treatment research: Lessons learned from the aphasia literature. Neuropsychology Review. 2006; 16:161–169. [PubMed: 17151940]

Busk, PL.; Serlin, RC. Meta-analysis for single-case research. In: Kratochwill, TR.; Levin, JR., editors. Single-case research design and analysis. Hillsdale, NJ: Lawrence Erlbaum; 1992. p. 187-212.

Campbell, DT.; Stanley, JC. Experimental and quasi-experimental designs for research. Boston, MA: Houghton Mifflin Co; 1963.

Campbell JM. Statistical comparison of four effect sizes for single-subject designs. Behavior Modification. 2004; 28:234–246. [PubMed: 14997950]

Cohen, J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum; 1988.

Durlak J. How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology. 2009; 34:917–928. [PubMed: 19223279]

Gierut JA, Elbert M, Dinnsen DA. A functional analysis of phonological knowledge and generalization learning in misarticulating children. Journal of Speech and Hearing Research. 1987; 30:462–479. [PubMed: 3695441]

Gierut JA, Morrisette ML. Age-of-word acquisition effects in treatment of children with phonological delays. Applied Psycholinguistics. 2011

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

McReynolds, LV.; Kearns, KP. Single-subject experimental designs in communicative disorders. Baltimore, MD: University Park Press; 1983.

Morrisette ML, Gierut JA. Lexical organization and phonological change in treatment. Journal of Speech, Language, and Hearing Research. 2002; 45:143–159.

Olive ML, Smith BW. Effect size calculations and single subject designs. Educational Psychology. 2005; 25:313–324.

Powell TW, Elbert M, Miccio AW, Strike-Roussos C, Brasseur J. Facilitating [s] production in young children: An experimental evaluation of motoric and conceptual treatment approaches. Clinical Linguistics & Phonetics. 1998; 12:127–146. [PubMed: 21434786]

Scruggs TE, Mastropieri MA, Casto G. The quantitative synthesis of single-subject research. Remedial and Special Education. 1987; 8:24–33.

Stoel-Gammon C. Relationships between lexical and phonological development in young children. Journal of Child Language. 2011; 38:1–34. [PubMed: 20950495]

Table I

Standard mean difference computations for Morrisette and Gierut (2002).

Condition	Child	M_A	M_B	SD_A[a]	d	d by condition
High frequency	1	1.212	27.241	1.664	16.07	12.6
	2	1.235	16.031	0.979	9.13	
Low frequency	3	4.348	12.169	1.537	4.83	5.9
	4	2.222	13.514	3.396	6.97	
Dense neighbourhoods	5	1.322	4.745	1.813	2.11	2.6
	6	2.247	7.258	2.185	3.09	
Sparse neighbourhoods	7	2.646	11.290	0.728	5.34	4.3
	8	3.333	8.621	0.661	3.26	

Notes: $$d = \frac{(M_B - M_A)}{SD_{A-\text{pooled across Ss}}}$$

[a]Each child's baseline *SD* is reported. The value of the pooled *SD*s (i.e. $SD_{A-\text{pooled across Ss}}$) = 1.620.