The Psychological Record, 1999, 49, 49-66

IMPROVING DESCRIPTIONS OF SINGLE-SUBJECT EXPERIMENTS IN RESEARCH TEXTS WRITTEN FOR UNDERGRADUATES

MARSHALL L. DERMER
University of Wisconsin—Milwaukee

THEODORE A. HOCH
Northern Virginia Training Center

Many undergraduate research methods texts assume that single-subject experiments are (a) low in generality, (b) merely pilot studies, (c) unsuitable when treatment effects are gradual or irreversible, (d) only useful when multiple-treatment interference is low, and (e) questionable because tests of statistical significance are absent. We critically examine these misconceptions, with the goal of improving the single-subject experiment's description in these texts and increasing its use beyond the area of behavior analysis. We subsequently compare multi-subject and single-subject experimentation in terms of research process and the alignment of psychological research with theory and practice. We conclude that psychological science and practice will be enhanced by methods texts accurately describing single-subject experiments and these texts addressing the problem of "individual-subject validity": the extent causal relations between treatments and outcomes are assessed at the level of the single-subject.

Science involves searching for relations among aspects of the universe and aggregating data across units of observation. Researchers in psychology have traditionally aggregated data across subjects and, within the past 15 years, they have begun to aggregate data across experiments (Cooper & Hedges, 1994). Psychologists aggregate data to cope with variability. Ironically, most psychological theory refers to the individual subject and it is not clear how relations derived by aggregating data across experiments or by aggregating data across subjects (Bass, 1987; Valsiner, 1986a, 1986b) can always be validly generalized to individuals.

Over the years, many types of validity have been discussed. Most research methods texts, however, have ignored "individual-subject validity": the extent causal relations between treatments and outcomes are assessed at the level of the individual subject. Individual-subject validity has concerned basic researchers in neuropsychology (Dywan &

Correspondence concerning this article should be addressed to Marshall Lev Dermer, Department of Psychology, University of Wisconsin-Milwaukee, Milwaukee, WI 53201. Electronic mail may be sent via Internet to dermer@uwm.edu.

Special acknowledgments are due Alan Baron, the late Donald T. Campbell, Ray Fleming, Susan Lima, John Surber, and Jay Moore for reviewing earlier versions of this manuscript.

Segalowitz, 1986, p. 296), practitioners in reading instruction (McCoy & Pany, 1986, p. 553) and others (Barlow, Hayes, & Nelson, 1984; Gelb, 1997) who are concerned about understanding and generalizing to the individual case. Individual-subject validity is, of course, a defining feature of single-subject experiments.

In single-subject experiments researchers typically measure one or more classes of performance of one or more subjects, over extended temporal intervals, systematically introduce and remove treatments, and visually inspect graphed data to determine whether and how these treatments controlled the performance of individual subjects. Worth noting is that single-subject experiments can include multiple subjects.

Given the obvious relevance of assessing causal relations, at the level of the individual subject, to psychological theory and applied psychology, it is surprising that single-subject experiments appear underutilized (Aeschleman, 1991) and are often misdescribed in research methods texts written for undergraduates (Kestner & Flora, 1995). Perhaps the apparent underutilization and misdescription result from single-subject methodology texts having been written by and for behaviorists (e.g., Barlow & Hersen, 1984; Johnston & Pennypacker, 1993; Kazdin, 1982; Poling & Fuqua, 1986; Poling, Methot, & LeSage, 1995; Sidman, 1960). Additionally, most undergraduate research texts appear to have been written by authors who have not conducted single-subject experiments.

Our objectives are to help such authors accurately describe single-subject experiments as well as encourage students and researchers to use these experiments in areas beyond behavior analysis. To achieve this, we critically review six misconceptions primarily drawn from undergraduate research methods texts.

Single-Subject Experiments Have Low Generality

Campbell and Stanley (1963, p. 5), in their sage and highly influential text, used the term internal validity to refer to the extent that changes in an experiment's treatments clearly cause changes in an experiment's outcome variables. When such a causal relation can be generalized across subject populations, settings, treatment variables, and outcome variables, they called the relation high in external validity. It is important to note that Campbell and Stanley neither explicitly considered individual-subject validity nor the related problem of generalizing to particular subjects. Despite these conceptual limitations, we may ask: How have undergraduate methods texts depicted the generality of single-subject experiments?

One methodology text admonishes its readers, "REMEMBER single-subject designs provide substantial internal validity, but have limited external validity" (Heiman, 1995, p. 330; also Graziano & Raulin, 1997, p. 313; Liebert & Liebert, 1995, p. 183). Another text goes further, "The results of the [single-subject] study can be generalized only to another identical organism in the same controlled setting" (Christensen, 1994, p.

384). If the organism and setting must be identical in every respect, is generalization ever possible?

Rather than emphasizing the low generality of single-subject experiments, other authors emphasize the higher generality of multi-subject experiments (e.g., Cozby, 1993, p. 101; Myers & Hansen, 1993, p. 212; 1997, p. 277). In these experiments researchers typically randomly assign subjects to two or more treatments, subsequently measure performance, and conduct tests of statistical significance to determine whether and how these treatments influenced performance aggregated across subjects. Many authors go further and assert that subjects and any other aspect of the experiment for which generalization is sought, for example the treatments, should be randomly sampled from the universe of generalization (McGuigan, 1997, pp. 361-367; Myers & Hansen, 1997, p. 189; Shaughnessy & Zechmeister, 1997, pp. 202-203).

In fact, however, such random sampling of subjects, treatments or other factors rarely occurs in psychological experiments. Instead researchers include a few treatments based on theoretical or applied considerations, attempt to set other factors to levels that will permit the treatments to be effective, and randomly assign readily available subjects to conditions. No doubt, there are populations in such research but they are finite and very selective. For example, a population of subjects might consist of all the introductory psychology students who showed up for Professor X's experiment, conducted on the fourth floor of Y Hall, during the third week of July, 1985, who could have been randomly assigned to the experimental treatment.

Although most undergraduate texts do recognize that researchers study such samples of convenience, these texts often do not properly acknowledge that researchers are, thereby, implicitly studying highly unique, hypothetical populations (see Campbell & Stanley, 1963, pp. 23-24; Winch & Campbell, 1969). Although these texts often cite Campbell and Stanley's (1963) methods text and refer to "external validity," few texts recognize that Campbell and Stanley *did not* advocate establishing generality by representatively sampling subjects, treatments, contextual operations, or dependent measures from some domain of generalization (p. 18; also see Jones, 1985, p. 67; Sidman, 1960, pp. 243-244).

Historically, scientists have discovered general, causal relations by carefully teasing apart those aspects of the universe that are relevant to the relation from those that are irrelevant. Campbell (1969) nicely illustrated this process with William Nicholson and Anthony Carlisle's discovery of electrolysis. Writing initially from the vantage point of a random sampling approach, Campbell noted:

> Taking in May, 1800, a very parochial and idiochronic sample of Soho water, inserting into it a very biased sample of copper wire, into which flowed a very local electrical current, they obtained hydrogen gas at one electrode, oxygen at the other, and uninhibitedly generalized to all the water in the world for all

> eternity. It was a hypothetical generalization to be sure, rather than a proven fact. There have been now many studies of the effects of "impurities" in the water upon [elect]rolysis, but these too have been done on very biased samples. The idea of a representative sampling of all the waters of the world . . . never occurred as an ideal. The very concept of "impurities," of segregating the contents of water into "pure" stuff and the alien contents, is one which would never have emerged had a representative sampling approach been employed. In the successful sciences, generalizations have . . . emerged from checking in nonrepresentative ways on an initial bold generalization. Scientists assumed that [elect]rolysis held true universally until it was shown otherwise. (pp. 360-361)

In other words, generalization is the product of careful analysis in which scientists first identify a causal relation and then probe its validity by examining if the relation can be replicated across experiments, particularly those that are operationally different but presumably theoretically identical. Failures to replicate indicate that the relation is not fully understood. Through a series of such theory-guided replications, scientists can gradually delimit the domain of generality (Lykken, 1968; Sidman, 1960, Chaps. 3-4).

This aspect of multi-subject research also characterizes the single-subject experiment as illustrated by Johnston and Pennypacker's (1980) comments about research in reading.

> Instead of asking how many children in the 6-year-old population will learn to read to a certain criterion using a new reading skills program, we should be asking what prerequisite skills are necessary for program effectiveness with every child and what variations in style of management are necessary under what conditions. In other words, an understanding of the generality of a procedure comes not from blindly testing larger and larger proportions of the population, but from a thorough understanding of the variables controlling its effec[t]s under any circumstances. (p. 401)

A researcher can discover the critical variables operating at the level of the individual subject through a series of single-subject experiments that utilize operationally different but theoretically identical procedures (Birnbrauer, 1981; Johnston & Pennypacker, 1993, pp. 244-253, 348-357; Sidman, 1960, pp. 110-139, 243-244). Thorngate (1986) similarly remarked in discussing cognitive processes, that "to find out what people do in general, we must first discover what each person does in particular, then determine what, if anything, these particulars have in common" (p. 75).

It is important to note that many authors, as suggested by the title of this section, refer to the generality of an experiment. An experiment, however, is necessarily bound in space and time and so is without generality. Rather, it is *aspects* of an experiment that may have

generality. These aspects, properly interpreted, are the groundwork for our theories.

We may, therefore, ask, "Are the relationships isolated through single-subject research less general than those isolated through multi-subject research?" This is an empirical question that has not been addressed. It should be noted, however, that when we can isolate the critical variables controlling behavior for particular persons, through single-subject research, then we can specify the treatments controlling behavior for large classes of individuals. This point is well illustrated by the promising work of Lovaas (1993) and his associates with autistic children. It now appears that a substantial proportion of such children will not have an autistic diagnosis if they receive about two years of intensive, behaviorally based instruction, beginning at about their second year (Lovaas, Calouri, & Jada, 1989; McEachin, Smith, & Lovaas, 1993).

Single-Subject Experiments Are Merely Pilot Studies

There is no generally accepted definition of "pilot" research, but apparently such research is preliminary either to determining whether a treatment can be implemented or to forecasting whether a full-scale experiment will be successful. In this regard, how have methodology texts depicted single-subject experiments?

Campbell and Stanley (1963, pp. 43-45) classify single-subject designs as "quasi-experimental designs"; by definition, such designs lack full experimental control (p. 34). The epithet "quasi-experimental" is also used by Rosenthal and Rosnow (1991) who note that "none is unambiguous in terms of internal validity (p. 97)." Levin and Hinrichs (1995) note that "extraordinary efforts must often be taken in order to infer causality in single subject experiments" (p. 36).

Others note that single-subject designs only make good pilot studies whose findings should be replicated with multi-subject experiments (e.g., Christensen, 1994, pp. 384; Myers & Hansen, 1993, p. 210). For those interested in individuals, however, the converse is true. For example, in discussing reading research, which almost exclusively uses multi-subject experiments, McCoy and Pany (1986) wrote:

> Teachers must collect performance data on effects of specified instructional techniques (such as corrective feedback) on individual students' learning. Analyzing their own data, teachers can confidently determine which [multi-subject] research findings and recommendations apply to their students. (p. 553)

An adequate research method should at least produce replicable results. *The Journal of the Experimental Analysis of Behavior* and *The Journal of Applied Behavior Analysis* most often publish reports of single-subject experiments. If these experiments were inherently flawed then there would be abundant failures to replicate and reports of a crisis in behavior analysis. Instead, single-subject experiments have contributed to a powerful behavioral technology (Meehl, 1978).

Single-Subject Experiments Are Unsuitable When Treatment Effects Are Gradual or Irreversible

Here, we briefly discuss two separate warnings: Avoid single-subject experiments when treatment effects are either gradual or irreversible (e.g., Cozby, 1993, p. 101; Liebert & Liebert, 1995, p. 181).

Single-subject experiments are appropriate when treatments produce gradual change. In this circumstance, the treatment may be kept in effect until the outcome measure does not vary considerably among observational periods. In a laboratory experiment, for example, this may mean that the last five sessions of data are only judged stable if a trend is visually absent and the individual observations do not differ from the five-session mean by more than a given percentage. Once the stability criterion is satisfied and a "steady state" is achieved, a comparison condition can be implemented. When stability is achieved for the comparison condition, the original condition may be reinstated. Treatments are compared *only* during the steady state. Such steady state analyses may involve many *intra-subject* treatment alternations to verify that the treatments reliably controlled behavior.

Steady state analyses can be difficult. There is, for example, no rule for choosing stability criteria. Generally, one must examine other researchers' successes and failures. Moreover, it is possible never to achieve a steady state either because the stability criterion is too stringent or the researcher has not exercised sufficient experimental control (see Perone's, 1991, excellent discussion of stability criteria). Nevertheless, there are research areas where outcome measures can be particularly stable and the major problem is finding an effective treatment, as illustrated by Chadwick and Lowe's (1990) work on delusional beliefs which often appear unchangeable.

And what about irreversibility? In this case, the multiple baseline design is most appropriate. This design utilizes three or more independent classes of behavior (e.g., spelling correctly, solving math problems, knowing physiological terminology) that, in principle, are each susceptible to treatment (e.g., tutoring). Once measures of these behaviors are stable, the treatment is applied to one behavior with the remaining behaviors serving as controls. When the change for the treated behavior stabilizes, the treatment is next applied to the second behavior and so on. This design is discussed in most single-subject methodology texts (e.g., Barlow & Hersen, 1984). In some versions of this design, multiple subjects must be used as, for example, in physiological research where a treatment requires ablating a cerebral structure. It must be noted again, however, that even in this case the design's purpose is to assess causation at the level of the individual subject.

Single-Subject Experiments Can Only Be Used When Multiple-Treatment Interference Is Low

Sequence, order, and carryover effects all refer to a treatment's effect depending on an earlier treatment (Hains & Baer, 1989). This, of

course, is also a potential problem in the traditional within-subjects designs used by multi-subject researchers. Is multiple-treatment interference a particularly serious threat to single-subject experiments as some methods texts suggest (e.g., Heiman, 1995, p. 330; Kamil, 1984, p. 49)?

Surely, multiple-treatment interference appears to be a cogent threat because psychological processes are largely contextual. An organism whose responses to one treatment could not depend on other treatments previously in effect (or other treatments concurrently in effect) would be at a great disadvantage. However, multiple-treatment interactions may not always be serious.

Consider the multielement, single-subject design (also called the alternating treatments design). In this design, two or more treatments are rapidly changed within some relatively short period (e.g., an hour or a day) and behavior is assessed during or after the treatments. If treatment effects always depended heavily on context then treatments, so administered, would not often produce differential performances. Such performances, however, are reported in the many experiments that have used this design.

In applied research, the "best" treatment may subsequently be administered alone to assess its effect when isolated from the remaining treatments. To what extent is the level of performance for the treatment during this isolation phase equivalent to the level of performance when the treatment was presented during the multielement phase? Hains and Baer (1989) reviewed the multielement designs presented in Barlow and Hersen's (1984) text and reported that for only four of fourteen such comparisons was performance equivalent across isolation and alternation phases. This result, of course, indicates that treatment effects can depend on context. How shall we address these and other potential interaction effects?

One approach is to administer treatments in random or counterbalanced orders and then aggregate performances for each treatment across orders. Such approaches, however, do not eliminate interference should it be present. Counterbalancing orders, however, does have an advantage. If each order is sufficiently replicated then one can examine whether a treatment's effect varies across orders.

A second approach is to reduce carry-over effects. In the multi-element design these effects may be attenuated by making treatments discriminable and by spacing them properly (McGonigle, Rojahn, Dixon, & Strain, 1987; Perone, 1991, p. 160). Spacing treatments also makes sense for other single-subject designs where each treatment is in effect for a long period, for example, a week. In such designs, spacing occurs, in effect, when treatments are only compared in their steady states or when a common baseline condition is interposed between treatments (Perone, 1991, p. 156).

A carry-over effect of particular concern to single-subject researchers is the consequence of activities before an experimental session on behavior during an experimental session. Multi-subject

researchers may often ignore whether one subject napped or another played an hour of handball before participation; single-subject researchers, in contrast, often attempt to control for such activities by disregarding performances from the early "warm-up" portions of sessions.

A third approach to multiple-treatment interference is to relabel it "multiple-treatment interaction" and study it (Sidman, 1960, pp. 334-335). As discussed earlier, knowledge of causal variables is critical for establishing generality.

Sidman discussed two methods for studying interactions using the multielement design. In using *independent verification,* a treatment is studied alone and in combination (either concurrently or in rapid alternation) with other treatments. In using *functional manipulation* some parameter of one or more treatments is systematically altered and combinations of the resulting treatments are studied, including factorial arrangements. Hains and Baer (1989) nicely provide examples of both approaches. In illustrating *functional manipulation* they present a series of single-subject designs and actual experiments that correspond to the ubiquitous multi-subject factorial design so that such interactions can be studied at the level of the single-subject. The authors of undergraduate research methods texts who consider the single-subject experiment "not well suited" for studying such interactions (e.g., Leary, 1995, p. 304) can review their report as well as the *Journal of the Experimental Analysis of Behavior* where such designs are often used although not described as factorial (Perone, 1991, p. 161).

Statistical Significance Tests Are Not Used in Single-Subject Research

Goodwin's (1995) methodology text correctly recognizes that:

> Single-subject designs are also criticized for not using statistical analyses but relying instead on a simple visual inspection of the data. To some extent, this reflects a philosophical difference between those advocating large and small N. Defenders of small N argue that conclusions are only drawn when the effects are large enough to be obvious to anyone. (p. 323)

Can the absence of statistical significance testing be addressed?

Suppose that in a single-subject experiment Treatment A is in effect, then B is in effect, and then A is reinstated. Suppose that the rate of behavior is assessed 10 times in each phase and that the data for the A phases are uniformly lower than the data for the B phase. In discussing the limitations of single-subject experiments, two editions of one methodology text ignored phase length and stability and instead focused on the two transitions reasoning:[1]

[1]Special thanks are extended to Alan Baron for bringing this example to our attention and having suggested, some ten years ago, in a colloquium, this essay's theme.

> This means that a good ABA reversal shows only a change from A_1 to B and an opposite change from B to A_2. The probability of each of these directional changes is 0.50, so the likelihood of obtaining the expected pattern is 0.50 X 0.50 or 0.25. It would, therefore, occur by chance in one-quarter of all cases in the absence of a real experimental effect. (Liebert & Liebert, 1995, p. 183)

But the basis for inferring a treatment effect is not merely the two transitions but the entire pattern of data involving temporally extended and consistently higher responding during Treatment B than during Treatment A. If one ignores "ties," then although it is true that one data point can either be above or below another, it is not true that the probability of a directional change is 0.5 given a baseline of 10 low points. Put another way, suppose a patient who had been in a coma for 10 years awakened after a drug was administered. Is the probability of this transition in the absence of the drug 0.5, given the 10 years of coma?

Although Campbell and Stanley (1963) considered statistical significance testing necessary in the social sciences, they also recognized it was unnecessary in other sciences.

> If the more advanced sciences use tests of significance less than do psychology and education, it is undoubtedly because the magnitude and the clarity of the effects with which they deal are such as to render tests of significance unnecessary. If our conventional tests of significance were applied, high degrees of significance would be found. It seems typical of the ecology of the social sciences, however, that they *must* work the low-grade ore in which tests of significance are necessary. (italics added, pp. 42-43)

But could the ore be low-grade *because* researchers use statistical significance tests to control for variability rather than use experimental control (Michael, 1974)?

By definition, experimenters exert control. To the multi-subject researcher "control" refers to randomly assigning subjects to differing conditions, otherwise keeping these conditions constant, averaging data across subjects, and using statistical significance tests to compensate for within-condition variability. To the single-subject researcher, however, "control" refers to directly eliminating sources of variability, besides the treatments, until treatment effects are visually apparent for each subject (see Mook, 1974, pp. 237-238). Often this requires tailoring treatments to each subject and treatment condition (e.g., Wolery & Ezell, 1993). The clearest example of this is the stability criterion which often results in treatments being in effect, from condition-to-condition and from subject-to-subject, for variable numbers of sessions. A more parochial example comes from studies of choice. Often subjects enter the experiment preferring one alternative and must each be given special training before the experiment can begin. As noted earlier, in the section addressing the alleged low generality of single-subject experiments, such experimental control promotes generalization.

Such direct experimental control also makes tests of statistical significance unnecessary. In this regard, Meehl (1978) noted "that most so-called "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless" (p. 806). In the case of clinical practice, he identified five "noble" traditions that he believed would be with us in the next 50 to 100 years. He could only recognize one feature that these traditions shared with the developed sciences, "they were originally developed with negligible reliance on *statistical significance testing*" (p. 817).

Implications

Having considered the most frequent criticisms of single-subject experimentation, it is important to consider the consequences of these misdescriptions.

The authors of several of the texts, reviewed above, reported the number of copies sold of each edition. The reports ranged from 5,000 to 10,000 copies. If we focus on the lower estimate and assume that at least one half of these texts entered the used book market, then about 7,500 undergraduates, per edition, read, studied, and were examined on flawed descriptions of single-subject research. The small proportion of undergraduates earning advanced degrees is unlikely to be disabused of the resulting misconceptions. A survey of graduate psychology programs (Aiken, West, Sechrest, & Reno, 1990) revealed that only 10% of the department chairs believed that most of their graduates could competently conduct single-subject experiments. The misconceptions about and ignorance of single-subject experiments produce at least three serious untoward consequences.

The first untoward consequence of inadequate knowledge about single-subject experiments is that potential researchers are unaware of the marked differences between single-subject and multi-subject research processes.

Psychologists are proud of being data oriented, but multi-subject researchers are not highly motivated to frequently monitor data collection because monitoring is unlikely to reveal whether a treatment is effective. Also, once a multi-subject experiment is implemented it should not be altered. Perhaps the most exciting moments in multi-subject research occur when after weeks or months of data collection the computer generates the statistical analyses and the results are interpreted.

Single-subject researchers, in contrast, are more highly motivated to frequently monitor data collection because the experiment's design depends on the subject's performance. Performance is plotted as it is generated so each session can punish or reward researchers' activities. If baseline performance appears sufficiently stable, and if a new treatment is effective, then researchers can usually see over a few weeks, days, or, for some problems and designs, within one session, the beginning of a gratifying shift in performance which may stabilize and be replicable. If such

efforts fail, single-subject researchers must tinker with treatments, stability criteria, and so forth. Barlow et al. (1984), the authors of one of the best introductions to single-subject experimentation for clinical and educational scientist-practitioners, call this attitude "investigative play."

Such "play" can have serious consequences. For example, consider research on learning. A single-subject researcher who permitted subject motivation to vary from session to session could not easily establish a baseline and would seek further experimental control. The researcher could, for example, enhance stability by providing rewards for appropriate performance (e.g., Perone & Baron, 1983). In contrast, a multi-subject researcher could validly reason that although motivation may differ from subject to subject, within the limits of sampling, random assignment of subjects to groups equates motivation across groups. With sufficient statistical power, the multi-subject researcher could detect statistically significant differences yet ignore subject motivation. The multi-subject experiment would be "successful" even though an important determinant of performance is ignored[2]; a determinant a successful practitioner such as a teacher could rarely ignore. Clearly, depending on statistical procedures to control for sources of variability reduces concern for developing important experimental procedures and technologies (Michael, 1974).

The second untoward consequence of inadequate knowledge about single-subject experiments is perpetuating the misalignment between psychological research and theory.

Every methodology text should present illustrations, like Figure 1 (Sidman, 1960), that depict how simple functional relations discernable at the level of the single-subject can produce between-subject variability. The upper panel illustrates functions for four subjects whose performances are linearly related to the treatments. The lower panel illustrates functions for four subjects whose performances are curvilinearly related to treatments. Each subject's performances systematically vary with the treatments. In marked contrast, there is substantial between-subject variability for most treatments. Consider, for example, how level X of the independent variable produces variable behavioral measures among the subjects. Such variability has, in part, led researchers to aggregate data across subjects. Only under very specific circumstances, however, will functions derived from data aggregated across subjects represent functions at the level of the single-

[2]Consider the recent debate about racial differences in intelligence. Although IQ tests are administered under standard conditions these conditions do not ensure that respondents are equally motivated. Herrnstein and Murray (1994) discussed but then dismissed racial differences in motivation (pp. 282-284). Nowhere, however, did they advocate that motivation be set at a high level when measuring IQ. In contrast, Bradley-Johnson, Johnson, Shanahan, Rickert, and Tardona (1984) awarded black and white children, of low socioeconomic status, tokens exchangeable for rewards, immediately after each correct response on the WISC-R. Black children's IQ scores increased at least 10 points, on the average, between the immediate reward condition and either a control condition or a delayed reward condition. White children's scores, in contrast, were not substantially influenced.

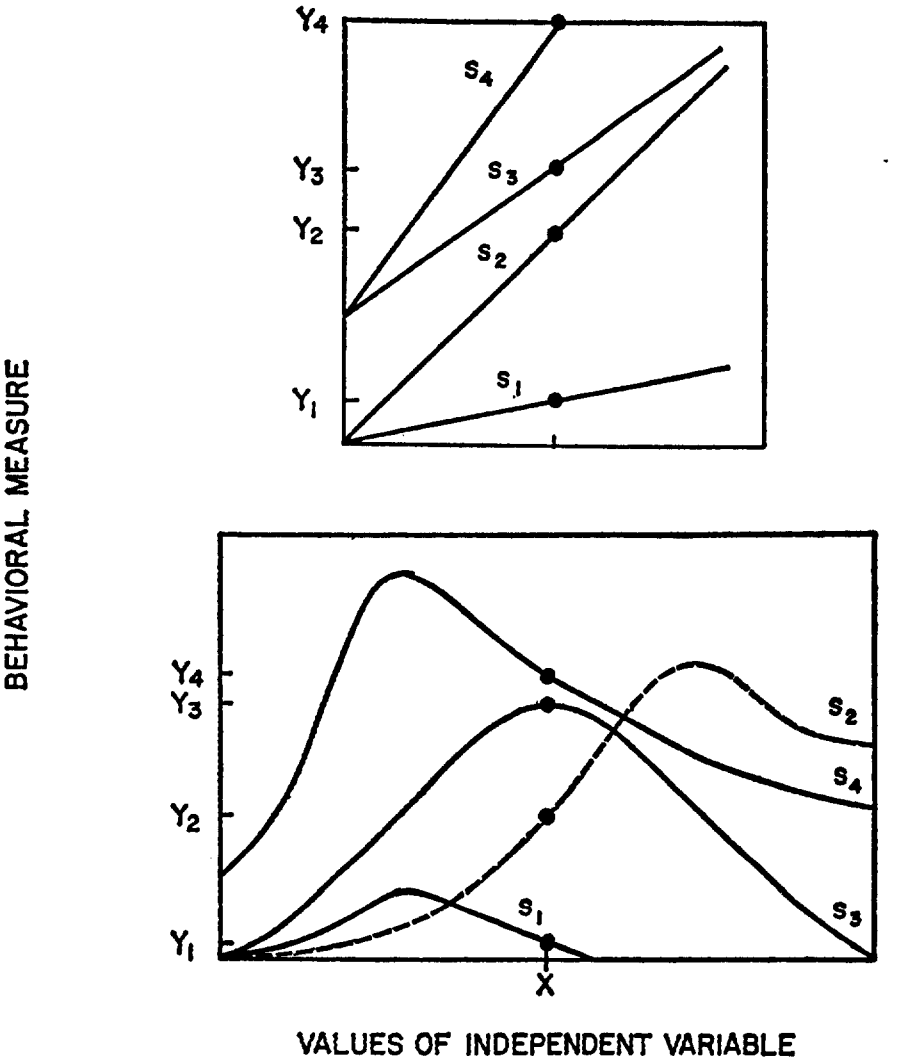

Figure 1. Upper panel: Functions for four subjects whose performances are linearly related to treatments. Lower panel: Functions for four subjects whose performances are curvilinearly related to treatments. Each subject's performance systematically varies with changes in the treatments. In marked contrast, there is substantial between-subject variability for most treatments. Consider, for example, how level X of the independent variable produces variable behavioral measures among the subjects.

Note. Panels from *Tactics of Scientific Research* (pp. 49-50), by Murray Sidman, 1960, New York: Basic Books. Copyright 1960 by Basic Books. Reprinted with permission.

subject (see Bass, 1987). Yet, it is the former functions on which most researchers base their theories![3]

There are, however, commentators who doubt that systematic relations can be observed at the level of the individual subject. Consider Lykken's (1991) "What's Wrong with Psychology Anyway?":

> Perhaps the only way to predict individual behavior with any precision or in any detail is idiographically, one individual at a time studied over months or years. To the extent that this is so, perhaps Psychology is really more like History than it is like Biology.

Lykken, not surprisingly, next suggests aggregating data:

> Maybe psychology is like statistical mechanics in the sense that we can make confident statements only about the means and variances of measurements on groups of people. (pp. 18-19)

And when aggregating data across subjects fails to produce order, we are now told to aggregate data across experiments.

Consider Schmidt's (1996) conclusions about what is needed to produce "cumulative knowledge in psychology" and the value of the individual experiment:

> Meta-analysis has explicated the critical role of sampling error, measurement error, and other artifacts in determining the observed findings and the statistical power of individual studies. In doing so, it has revealed how little information there typically is in any single study. It has shown that, contrary to widespread belief, a single primary study can rarely resolve an issue or answer a question. Any individual study must be considered a data point to be contributed to a future meta-analysis. Thus the scientific status and value of the individual study is necessarily lower than has typically been imagined in the past. (p. 127)

Meta-analysis is essentially conservative in accepting between-subject and between-experiment variability and detecting effects despite such variability. For Schmidt this means that:

> Cumulative understanding and progress in theory development is possible after all. It means that the behavioral and social sciences can attain the status of true sciences. (p. 123)

[3]Brown and Kirsner (1980) similarly illustrate how the correlation between variables each based on data aggregated across subjects need not describe the correlation between these variables at the level of the individual subject. They note, "The crucial point is not that the [correlations] represent different estimates of the strength of the relationship between a given pair of variables, but rather that they provide estimates of the strength of different relationships" (p. 184).

But what kind of science is this? If Schmidt is correct then we are developing a science, at least based on two levels of aggregation that is far removed from psychological theory which is essentially a theory about individuals. Future researchers need to understand that although aggregation is a valuable tool, the single-subject experiment often offers a better way than the multi-subject experiment of addressing variability and, perhaps, is the only practicable way of realigning psychological research with theory.

The third untoward consequence of inadequate knowledge about single-subject experiments is perpetuating the misalignment between psychological research and practice (Barlow et al., 1984). The true science that Schmidt envisions may help corporate presidents, deans, generals, high school principals, marketing professionals, and others who are responsible for aggregates but it will least help clinicians, counselors, parents, social workers, teachers, and others who are responsible for individuals.

Conclusion

Although behaviorists appear to have most often used single-subject experiments, the methodology is quite general. Many articles and monographs accurately describe single-subject experiments and advocate their greater use in areas such as audiological rehabilitation (Demorest & Erdman, 1994), clinical psychology and education (Barlow et al., 1984; Bryson-Brockman & Roll, 1996; Carlson, 1985; McCormick, 1990; Strain, McConnell, & Cordisco, 1983), counseling psychology (Tracey, 1983) and communicative disorders (Connell & Thompson, 1986; Kearns, 1986; McReynolds & Thompson, 1986). It is time for the authors of methods texts to improve their discussions of single-subject experiments.

Authors of methods texts, searching for illustrative experiments, can use "single-subject" and design names such as "multiple-baseline" to search the PsycLIT database. A recent search produced, for example, experiments on modifying delusional beliefs (Chadwick & Lowe, 1990; Himadi & Kaiser, 1992); and experiments on training sentence production which utilized, in part, work by Chomsky (Thompson, Shapiro, & Roberts, 1993; Thompson, Shapiro, Tait, Jacobs, & Schneider, 1996).

More generally, methods texts need to better discuss the benefits (Rushton, Brainerd, & Pressley, 1983) and costs (Sidman, 1960; Thorngate, 1986, pp. 73-76) of aggregation. Worth considering are Glass, McGaw, and Smith's (1981) closing remarks in their influential introduction to meta-analysis:

> 1. There is ". . . a fairly large residual unpredictability in *individual* human behavior."
> 2. "The condition of most social and behavioral research appears to be that there is little predictability at either the individual or study level."
> 3. "Where does the remedy lie for this affront to scientific aspirations? . . . We have no clear sense of a solution, nor even whether any solution is possible." (pp. 230-231)

The solution may reside, ironically, in minimizing data aggregation across subjects and across experiments and maximizing experimental control at the level of the single subject. To achieve this, research methods texts must better describe single-subject experiments and address the problem of individual-subject validity.

References

AESCHLEMAN, S. R. (1991). Single-subject research designs: Some misconceptions. *Rehabilitation Psychology*, 36, 43-49.

AIKEN, L. S., WEST, S. G., SECHREST, L., & RENO, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, 45, 721-734.

BARLOW, D. H., HAYES, S. C., & NELSON, R. O. (1984). *The scientist practitioner: Research and accountability in clinical and educational settings*. New York: Pergamon.

BARLOW, D. H., & HERSEN, M. (1984). *Single case experimental designs* (2nd ed.). Elmsford, NY: Pergamon.

BASS, R. F. (1987). The generality, analysis, and assessment of single-subject data. *Psychology in the Schools*, 24, 97-104.

BIRNBRAUER, J. S. (1981). *External validity and experimental investigation of individual behavior*, 1, 117-132.

BRADLEY-JOHNSON, S., JOHNSON, C. M., SHANAHAN, R. H., RICKERT, V. I., & TARDONA, D. R. (1984). Effects of token reinforcement on WISC-R performance of black and white, low socioeconomic second graders. *Behavioral Assessment*, 6, 365-373.

BROWN, H. L., & KIRSNER, K. (1980). A within-subjects analysis of the relationship between memory span and processing rate in short-term memory. *Cognitive Psychology*, 12, 177-187.

BRYSON-BROCKMAN, W., & ROLL, D. (1996). Single-case experimental designs in medical education: An innovative research method. *Academic Medicine*, 71, 78-85.

CAMPBELL, D. T. (1969). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 351-382). New York: Academic Press.

CAMPBELL, D. T., & STANLEY, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

CARLSON, P. E. (1985). Updating and broadening the use of single subject designs in reading. *Reading Psychology*, 6, 251-265.

CHADWICK, P. D. J., & LOWE, C. F. (1990). Measurement and modification of delusional beliefs. *Journal of Consulting and Clinical Psychology*, 58, 225-232.

CHRISTENSEN, L. B. (1994). *Experimental methodology* (3rd ed.). Boston: Allyn and Bacon.

CONNELL, P. J., & THOMPSON, C. K. (1986). Flexibility of single-subject experimental designs. Part III: Using flexibility to design or modify experiments. *Journal of Speech and Hearing Disorders*, 51, 214-225.

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

COZBY, P. C. (1993). *Methods in behavioral research* (5th ed.). Mountain View, CA: Mayfield.

DEMOREST, M. E., & ERDMAN, S. E. (1994). Research in audiological rehabilitation: The challenges. [Monograph Supplement]. *Journal of the Academy of Rehabilitative Audiology*, 27, 291-315.

DYWAN, J., & SEGALOWITZ, S. J. (1986). The role of the case study in neuropsychological research. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 291-310). New York: Plenum Press.

GELB, S. A. (1997). The problem of typological thinking in mental retardation. *Mental Retardation*, 35, 448-457.

GOODWIN, C. J. (1995). *Research in psychology methods and design*. New York: John Wiley.

GLASS, G. V., MCGAW, B., & SMITH, M. L. (1981) *Meta-analysis in social research*. Beverly Hills, CA: Sage.

GRAZIANO, A., & RAULIN, M. L. (1997). *Research methods/A process of inquiry*. (3rd ed.). New York: Longman.

HAINS, A. H., & BAER, D. M. (1989). Interaction effects in multielement designs: Inevitable, desirable, and ignorable. *Journal of Applied Behavior Analysis*, 22, 57-69.

HEIMAN, G. W. (1995). *Research methods in psychology*. Boston: Houghton Mifflin.

HERRNSTEIN, R. J., & MURRAY, C. (1994). *The bell curve*. New York: Free Press.

HIMADI, B., & KAISER, A. J. (1992). The modification of delusional beliefs: A single-subject evaluation. *Behavioral Residential Treatment*, 7, 1-14.

JOHNSTON, J. M., & PENNYPACKER, H. S. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Lawrence Erlbaum.

JOHNSTON, J. M., & PENNYPACKER, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

JONES, E. E. (1985). Major developments in social psychology during the past five decades. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. I, pp. 47-107). New York: Random House.

KAMIL, M. L. (1984). Current traditions of reading research. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 39-62). New York: Longman.

KAZDIN, A. E. (1982). *Single-case research designs: methods for clinical and applied settings*. New York: Oxford University Press.

KEARNS, K. P. (1986). Flexibility of single-subject experimental designs. Part II: Design selection and arrangement of experimental phases. *Journal of Speech and Hearing Disorders*, 51, 204-213.

KESTNER, J., & FLORA, S. R. (1995). Representation of behavioral methodology in experimental psychology textbooks. *The Behavior Analyst*, 18, 385-390.

LEARY, M. R. (1995). *Introduction to behavioral research methods* (2nd ed.). Belmont, CA: Wadsworth.

LEVIN, I. P., & HINRICHS, J. V. (1995). *Experimental Psychology/Contemporary methods & applications*. Madison, WI: WCB Brown & Benchmark.

LIEBERT, R. M., & LIEBERT, L. L. (1995). *Science and behavior* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

LOVAAS, O. I. (1993). The development of a treatment-research project for developmentally disabled and autistic children. *Journal of Applied Behavior Analysis*, 26, 617-630.

LOVAAS, I, CALOURI, K., & JADA, J. (1989). The nature of behavioral treatment and research with young autistic persons. In C. Gillberg (Ed.), *Diagnosis and treatment of autism* (pp. 285-305). New York: Plenum Press.

LYKKEN, D. T. (1968). Statistical significance testing in psychological research. *Psychological Bulletin*, 70, 151-159.

LYKKEN, D. T. (1991). What's wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology: Vol. 1: Matters of Public Interest* (pp. 3-39). Minneapolis: University of Minnesota Press.

MCCORMICK, S. (1990). A case for the single-subject methodology in reading research. *Journal of Research in Reading*, 13, 69-81.

MCCOY, K. M., & PANY, D. (1986). Summary and analysis of oral reading corrective feedback research. *The Reading Teacher*, 40, 548-554.

MCEACHIN, J. J., SMITH, T., & LOVAAS, I. O. (1993). Long-term outcomes for children with autism who received early intensive behavioral treatment. *American Journal on Mental Retardation*, 97, 359-372.

MCGONIGLE, J. J., ROJAHN, J., DIXON, J., & STRAIN, P. S. (1987). Multiple treatment interference in the alternating treatments design as a function of the intercomponent interval length. *Journal of Applied Behavior Analysis*, 20, 171-178.

MCGUIGAN, F. J. (1997). *Experimental psychology* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

MCREYNOLDS, L. V., & THOMPSON, C. K. (1986). Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. *Journal of Speech and Hearing Disorders*, 51, 194-203.

MEEHL, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.

MICHAEL, J. (1974). Statistical inference for individual organism research mixed blessing or curse? *Journal of the Experimental Analysis of Behavior*, 7, 647-653.

MOOK, D. G. (1974). *Psychological research, strategy and tactics*. New York: Harper & Row.

MYERS, A., & HANSEN, C. H. (1993). *Experimental psychology* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

MYERS, A., & HANSEN, C. H. (1997). *Experimental psychology* (4th ed.). Pacific Grove, CA: Brooks/Cole.

PERONE, M. (1991). Experimental design in the analysis of free-operant behavior. In I. H. Iversen & K. A. Lattal (Eds.), *Experimental analysis of behavior* (pp. 135-171). New York: Elsevier.

PERONE, M., & BARON, A. (1983). Reduced age differences in omission errors after prolonged exposure to response pacing contingencies. *Developmental Psychology*, 19, 915-923.

POLING, A., & FUQUA, R. W. (Eds.). (1986). *Research methods in applied behavior analysis*. New York: Plenum Press.

POLING, A., METHOT, L. L., & LESAGE, M. G. (1995). *Fundamentals of behavior analytic research*. New York: Plenum Press.

ROSENTHAL, R., & ROSNOW, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

RUSHTON, J. P., BRAINERD, C. J., & PRESSLEY, M. (1983). Behavioral development and construct validity. *Psychological Bulletin*, 94, 18-38.

SCHMIDT, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychological Methods*, 1, 115-129.

SIDMAN, M. (1960). *Tactics of scientific research*. New York: Basic Books.

SHAUGHNESSY, J. J., & ZECHMEISTER, E. B. (1997). *Research methods in psychology* (3rd ed.). New York: McGraw-Hill.

STRAIN, P. S., MCCONNELL, S., & CORDISCO, L. (1983). Special educators as single-subject researchers. *Exceptional Education Quarterly*, 4, 40-51.

THOMPSON, C. K., SHAPIRO, L. P., & ROBERTS, M. M. (1993). Treatment of sentence production deficits in aphasia: A linguistic-specific approach to *wh*-interrogative training and generalization. *Aphasiology*, 7, 111-113.

THOMPSON, C. K., SHAPIRO, L. P., TAIT, M. E., JACOBS, B. J., & SCHNEIDER, S. L. (1996). Training *wh*-question production in agrammatic aphasia: Analysis of argument and adjunct movement. *Brain and Language*, 52, 175-228.

THORNGATE, W. (1986). The production, detection, and explanation of behavioral patterns. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 71-93). New York: Plenum Press.

TRACEY, T. J. (1983). Single case research: An added tool for counselors and supervisors. *Counselor Education and Supervision*, 22, 185-196.

VALSINER, J. (Ed.). (1986a). *The individual subject and scientific psychology.* New York: Plenum Press.

VALSINER, J. (1986b). Between groups and individuals. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 113-151). New York: Plenum Press.

WINCH, R. F., & CAMPBELL, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, 4, 140-143.

WOLERY, M., & EZELL, H. K. (1993). Subject descriptions and single-subject research. *Journal of Learning Disabilities*, 26, 642-647.